DOCUMENT RESUME

ED 346 808 HE 025 628

TITLE	National Assessment of College Student Learning: Issues and Concerns. A Report on a Study Design Workshop.
INSTITUTION	National Center for Education Statistics (ED), Washington, DC.
REPORT NO	ISBN-0-16-037965-2; NCES-92-068
PUB DATE	Jun 92
NOTE	118p.; "With special reports by Addison Greenwood."
AVAILABLE FROM	U.S. Government Printing Office, Superintendent of Documents, Mail Stop: SSOP, Washington, DC 20402-9328.
PUB TYPE	Collected Works - Conference Proceedings (021)
EDRS PRICE	MF01/PC05 Plus Postage.
DESCRIPTORS	College Outcomes Assessment; College Students; Educational Assessment; *Educational Planning; Higher Education; *Outcomes of Education; Position Papers; Problem Solving; Seminars; *Student Development; *Student Evaluation; Thinking Skills; Workshops
IDENTIFIERS	*National Center for Education Statistics

ABSTRACT

This report presents the results of a workshop, held in Arlington, Virginia, on November 17-19, 1991, to discuss with the larger community the National Center for Education Statistics' (NCES) effort to develop strategies for assessing college student learning in support of National Education Goal Five, Objective Five which supports a substantial increase the proportion of college graduates who demonstrate advanced reasoning and communication skills. It is noted that of particular interest is the identification of the issues and concerns that NCES must consider in developing such an assessment process. The report begins with a brief description of the project goals and activities and is followed by a report of the workshop opening session and small group reports. Listed are the position papers by author and reviewers as well as general statements of the workshop participants. The general statements (individual comments) address what some participants would like to see as a user, what the most important next steps by NCES should be, what the major barriers and/or problems are that the NCES is likely to face, and who else should be consulted. (GLR)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

HE

ED346808

HE025628

NATIONAL CENTER FOR EDUCATION STATISTICS

National Assessment of College Student Learning: Issues and Concerns

A Report on a Study Design Workshop

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it

☐ Minor changes have been made to improve reproduction quality

- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

U.S. Department of Education
Office of Educational Research and Improvement
NCES 92-068

BEST COPY AVAILABLE

National Assessment of College Student Learning: Issues and Concerns

A Report on a Study Design Workshop

Project Officers:
Sal Corrallo
Gayle Fischer

with Special Reports by
Addison Greenwood

U.S. Department of Education
Lamar Alexander
Secretary

Office of Educational Research and Improvement
Diane Ravitch
Assistant Secretary

National Center for Education Statistics
Emerson J. Elliott
Acting Commissioner

National Center for Education Statistics

"The purpose of the Center shall be to collect, and analyze, and disseminate statistics and other data related to education in the United States and in other nations."—Section 406(b) of the General Education Provisions Act, as amended (20 U.S.C. 1221e–1).

June 1992

For sale by the U.S. Government Printing Office
Superintendent of Documents, Mail Stop: SSOP, Washington, DC 20402-9328
ISBN 0-16-037965-2

FOREWORD

In 1990, the National Education Goals Panel established long-term objectives to guide America toward educational excellence. National Education Goal 5 states that "By the year 2000, every adult American will be literate and possess the knowledge and skills necessary in a global economy and exercise the rights and responsibilities of citizenship." Five objectives are listed under the goal, one of which is directed at college student learning. It reads, "The proportion of college graduates who demonstrate an advanced ability to think critically, communicate effectively, and solve problems will increase substantially." In order to track student progress toward fulfilling this objective, a strategy for assessing these skills must be identified.

In response to this need, the National Center for Education Statistics hosted a study design workshop in the Fall 1991. In preparation for that workshop, a selected group of researchers, policymakers, and practitioners were invited to prepare "position papers" presenting their viewpoints on the issues and providing supporting documentation for their stance. Each paper was reviewed by three persons involved in some aspect of college student learning and assessment. The position papers and reviewers established the agenda for the workshop that followed. Eighty researchers, practitioners and policymakers (including the authors and reviewers) having longstanding interest and expertise in the area of college student learning and assessment gathered in Arlington, Virginia on November 17-19, 1991.

The workshop along with the commissioned papers and reviews provide valuable suggestions and insights for the design of an assessment process. Although many of the publics concerned with the assessment of higher learning have been included in this activity, others remain to be consulted, and thus the initial task is far from complete.

This publication, and the activities from which it came, would not have been possible without the work and support of the authors, reviewers, and participants. They came from both inside and outside the Federal Government, and brought with them a wide range of experiences and interests in the assessment of student learning. That the development of a process to assess the higher order thinking and communication skills of college graduates is an important and monumental task is of little doubt. However the enthusiasm and support given by all involved was beyond expectations. The written and spoken contributions of the participants are well documented. Regrettably, it was more difficult to capture the seriousness of purpose and the open mindedness of spirit with which the participants faced the challenge. I wish to extend my personal thanks to all who made it possible.

Emerson J. Elliott
Acting Commissioner, NCES

TABLE OF CONTENTS

Foreword iii

Introduction 1

Background 3
Sal Corrallo & Gayle Fischer, NCES

Listing of Position Papers by Author with Reviewers 17

The Study Design Workshop (Compiled by Addison Greenwood)

Opening Session: Charge to Participants,
Emerson Elliott, Acting Commissioner, NCES 23

Small Group Leader Reports to All Participants 33

Small Group Synthesis Reports 55

General Statements, Workshop Participants 75

Closing Exercise 101

Closing Comments 115
Gary Phillips, Acting Associate Commissioner, EAD
Emerson Elliott, Acting Commissioner, NCES

INTRODUCTION

The purpose of the study design workshop was to begin discussion with the larger community regarding the NCES effort to develop strategies for assessing college student learning in support of National Education Goal Five, Objective Five. Of particular interest is identification of the issues and concerns that NCES must consider in developing such an assessment process. The results of that workshop are presented in detail in this report.

The publication begins with a brief description of the project goals and activities and is followed by a report of the conference proceedings. As noted in the Foreword, in preparation for the conference a number of position papers were commissioned along with scholarly reviews. They served as background information for participants. A complete listing of the papers and reviewers are included in the report (pages 20-22) along with information on how they may be obtained. They are rich in content and ideas. The listing also includes two valuable pieces provided voluntarily by conference participants, Peter Facione and Michael Scrivan.

BACKGROUND

The National Goals Resource Group March 1991 Interim Report noted that,

> "...neither national nor state information is currently available on the ability of college graduates to "think critically, communicate effectively, and solve problems."

The report also suggested that,

> "If the National Goals Panel wishes to assess the ability of college graduates to think critically, to communicate effectively, and to solve problems, a new kind of assessment will have to be created. That assessment might be a type of National Assessment of Educational Progress (NAEP) at the college level, given to a national sample of college students at different kinds of institutions across the Nation. To have credibility, such an assessment would have to take into account differences in the postsecondary institutions in America and the fact that the pluralistic system in place today has extended postsecondary educational opportunities to the broadest cross section ever of America's citizens. Developing a NAEP like assessment would be controversial for many reasons. It would require 5 years or more to develop and an investment of several scores of millions of dollars to make operational." [1]

Given this challenge, the workshop planning team concluded that the process of identifying one or more assessment strategies should begin with consideration of four primary concerns or issues. These are listed below:

> First, what specific skills would or should be affected by student cognitive and/or affective learning experiences, as they relate to critical thinking, communication and problem solving abilities? Considering the larger goal of developing a competitive workforce and enhancing citizenship skills, must there be a common understanding, and perhaps agreement, about what specific skills students should achieve? The objective under the goal specifically designates college graduates. Can the assessment of higher order thinking and communication skills, which apply to job function and citizenship, occur that close to the college experience? Can the achievement of these skills be considered an end unto themselves or must they be considered in the context of course content or the total academic learning experience? Once clarified,

---

[1] National Educational Goals Resources Group, Interim Report, March 1991.

can they also be defined in a manner to allow for assessing the impact of the educational experience? Should or could they be defined from a teaching/learning perspective, so that their enhancement can be factored into classroom experiences? What has been the experience of institutions and states with active assessment programs? [2]

Second, who is to be tested? The national goal focuses upon college graduates. However, some suggest that since college graduates are part of the larger age cohort, the assessment of skills for all in the age cohort must be considered. Specifically how do the higher order thinking and communication skills of those in the workplace, who did not attend college, compare with those who did attend a postsecondary institution? It is also important to recognize that college graduates come with various types of programs and degrees--two-year and four-year, graduate and professional school degrees--while some working will not have been in a formal educational program for several years. Can an approach be developed that samples an entire population and distinguishes skills attained in the postsecondary experience from those achieved in the process of maturation? It is noted that the Department of Labor Secretary's Commission on Achieving Necessary Work Skills (SCANS) is planning on using The Department of Education's National Adult Literacy Survey to develop a test which assesses work-related skills for young adults currently employed and out of school. A number of the skills identified are similar to those noted in Goal Five. Can the process developed through the SCANS project for young working adults also be used to assess the skills of recent college graduates as defined by the national goal? Such an approach would have to consider definitions, standards and instrumentation used for the diverse universe designated.

---

[2] A 1987 report provided suggestions for the State of New Jersey's College Outcomes Program. It used much the same language when referring to the "...broad based common skills that are necessary in all disciplines and fields." More specifically they "...include analysis, problem solving, critical thinking, quantitative reasoning, and written and oral expression. These skills are seen in a students ability to find, use, and present information." Further elaboration follows:

> "These are skills necessary to critically analyze and utilize information (sometimes referred to as "higher order" skills). Specifically they include the skills necessary to:
>
> a) Accumulate and Examine Information - including the skills necessary to: determine the kinds of information needed for a given task; construct and implement a systematic search procedure, using both traditional and computerized methods; discard or retain information based on initial screening for relevance and credibility; and develop abstract concepts appropriate to the task at hand for initially ordering information which is retained.
>
> b) Reconfigure, Think About, and Draw Conclusions from Information - including skills necessary to: evaluate the interpretations presented by others in terms of their assumptions, logical inferences, and empirical evidence; reconfigure information in ways that suggest ranges of alternative interpretations and evaluate their relative merits; construct hypotheses that logically extend thought from areas in which information is already available into areas where it is not; specify the additional information which might confirm or disconfirm those hypotheses; and draw conclusions based on all of the above.
>
> c) Present Information - Including the skills necessary to express one's own ideas in written, oral, and graphic forms which will be intelligible and persuasive to a variety of audiences."

Third, there is the question of performance standards. Low standards may reduce the value of the program, while high standards can be troublesome and perhaps unrealistic for both students and institutions. Additionally, all do not enter postsecondary education or the workplace with the same learning experiences, family background, or cognitive abilities. To expect all to have attained the same level of thinking and communication skills may not be realistic. In recognition of these differences, it is not uncommon for varying levels of proficiency to be defined for a skill area. Examples of such efforts include the New York State Education Department's "Basic and Expanded Basic Skills" study and the Fort Worth Independent School District's "Levels of Proficiency" scales . Should and how can levels of proficiency be defined, validated, and measured?

Fourth, what type of instrumentation and approaches could/should be used? Can such assessments maintain both validity and reliability over time? When and how often should the assessment be conducted? Is it necessary to monitor the progress of all or some of the those to be assessed and for how long? Are reliable assessment approaches or instrumentations available in the open market? Some have suggested that, in developing an assessment strategy, attention should be focused on assessing students through indirect, curriculum-based approaches at both the institutional and state level. In lieu of collecting information directly from respondents or creating new assessment procedures, are proxy measures available, at least for short term use? Is the information needed available in current data banks? Can new or existing information be collected directly from the workplace? Who will do the assessment--institutions, states, private agencies, the Federal Government? How much will an assessment program cost and who will pay? What problems are to be expected in a large scale testing program?

Secondary Concerns: In addition to the primary issues and concerns noted above, several related areas need to be addressed.

First, questions have been raised as to whether the focus should be limited only to identifying a process(es) for assessing the higher order thinking and communication skills, noted in Goal Five. Paul R. Pintrich, at a U.S. Department of Education-sponsored conference on postsecondary assessment in 1986, argued that the most effective assessment of learning takes place when a theory or model of how instruction will lead to critical thinking or problem solving has been defined and tested. "Not only will this allow for a more accurate assessment of learning, but equally important it would help to delineate how the independent variables of course tasks and activities, curriculum offerings, and/or institutional dimensions theoretically influence the dependent variables of students' critical thinking." [3] Can an assessment process be developed and implemented that also considers the effectiveness of the teaching/learning process within the time and resource limits? Alternatively, could a method be defined, at a later date, for assessing effective teaching practices which builds upon the measurement process developed for outcomes alone?

Second, some suggest that the questions identified under the Issues/Concerns section must be considered in sequence. Although it seems reasonable that until there is closure on what is to be assessed, research on performance standards or the assessment process must wait. However this is not a necessary condition for this project. The thrust of this project is exploratory. The purpose is to review the past and to gain an understanding of the present from the living experiences of individuals, institutions and states currently involved in the assessment of student learning. It is important that the totality of the experiences be identified and documented as it relates to a national effort. In some instances, however, the work may focus on single sets of questions. For example, in those cases where institutions and/or states have assessment programs, the focus of the paper may be upon measurement and implementation issues.

Third, a number of states and institutions are currently considering many of the same questions. Some interaction between state and institutional personnel is expected during the feasibility stage (suggestions will be solicited by those commissioned to write position papers). A more complete review of these activities is expected during the development of the assessment process.

Fourth, another concern relates to test objectives and delineation of specific skills to be measured. Some are concerned that a test administered at the national level would result in a *de facto* national program or standards that would affect course content and testing methods. The pros and cons on the issue are many and will need to be discussed in some detail as the development of an assessment process moves forward.

---

[3] Pintrich, Paul R., "Assessing Student progress In College: A Process-Oriented Approach to Assessment of Student Learning In Postsecondary Settings" in Postsecondary Assessment Conference: Report of the Planning Committee, November 20, 1986 U.S. Department of Education, Washington, D.C.

Study Design Workshop: Authors, reviewers, practitioners, and policymakers from institutions, states, and national organizations participated in the study design workshop held November 17-19, 1991, in Arlington, Virginia. The original papers, a summary document of the papers, and the comments of the reviewers served as background material. Participants separated into four small work groups supported by a leader (NCES senior staff) and a recorder. Groups, designated by the participants' areas of expertise/interest, addressed the same agenda from their different perspectives. Summary reports of the individual workgroups were presented during a full session. At the close of the workshop, all participants were given the opportunity to address four questions in a handwritten exercise: (1) What would you like to see as a user? (2) What are the most important next step(s) NCES should take? (3) What are the major barriers and/or problems we are likely to face? and (4) Who else should be consulted?

The information gathered from the papers, reviews, and workshop will be used to determine the next steps in the development of one or more methods for assessing the noted thinking and communication skills.

Closing Comments: Given the nascent level of this activity to develop strategies for assessing higher order thinking and communication skills of college graduates, authors were entrusted with a great deal of latitude and discretion in developing their positions. As such, the papers and reviews were short on providing specific approaches and long on eliciting the issues and concerns that require thought and attention. The workshop proceedings provide more definitive suggestions for addressing the four basic questions outlined in the "Primary Issues and Concerns" section.

Although the project was designed to assist in the development of a process to assess student learning, equally important is the potential use of the study results for improving the educational experience. Specifically, once the skills are identified, clarified and assessed, findings can and should have a profound impact on what is taught and how it is taught.

Sal Corrallo,
Workshop Coordinator

Gayle Fischer,
Assistant Workshop Coordinator

List of Workshop Participants and Group Leaders:

WORKGROUP #1

Leader:

Gary Phillips
NCES, Acting Associate Commissioner, Education Assessment Division

Participants:

Cliff Adelman
OERI, Office of Research

Lorenz Boehm
Oakton Community College, The Critical Literacy Project

Robert Calfee
Stanford University, School of Education

John Chaffee
LaGuardia Community College, Creative and Critical Thinking Studies

John Daly
University of Texas at Austin, Department of Communication Studies

Robert Ennis
University of Illinois, Urbana

Peter Facione
Santa Clara University, School of Arts and Sciences

Ron Hambleton
University of Massachusetts

Donald Lazere
California Polytechnic, English Department

Barbara Lieb
OERI, Programs for the Improvement of Practice

Gerald Nosich
Sonoma State University, Center for Critical Thinking

Susan Nummedal
California State University at Long Beach, Critical Thinking Project

Richard Paul
Sonoma State University, Center for Critical Thinking

Rebecca Rubin
Kent State University, School of Communication Studies

Mark Weinstein
Montclair State University, Institute for Critical Thinking

Pat Dabbs, Assisting Recorder, NCES

Sheida White, Assisting Recorder, NCES

Monica Schnell, University of Maryland, Recorder

WORKGROUP #2

Leader:

Ron Hall
NCES, Acting Associate Commissioner, Postsecondary Education Statistics Division

Participants:

Trudy Banta
University of Tennessee at Knoxville, Center for Assessment Research and Development

Ernst Benjamin
American Association of University Professors

Pat Courts
State University of New York at Fredonia, English Department

Peter Ewell
National Center for Higher Education Management Systems

Steve Gorman
NCES, Education Assessment Division

Michael Knight
Kean College

Charles Lenth
State Higher Education Officers Organization

Georgine Loacker
Alverno College, English Department

Ted Marchese
American Association of Higher Education

Jean McDonald
National Governors Association

Margaret Miller
Virginia State Council of Higher Education

Ed Morante
College of the Desert, School of Resources, Research and Technology

James Ratcliff
Pennsylvania State University, National Center for Postsecondary Teaching, Learning and Assessment

Jerry Sroufe
American Education Research Association

Ron Swanson
Texas Academic Skills Program

Susan Towombly
University of Kansas

Jeff Gilmore, Assisting Recorder, OERI, Office of Research

Chris Carr, University of Maryland, Recorder

WORKGROUP #3

Leader:

Jeanne Griffith
NCES, Associate Commissioner, Data Development Division

Participants:

Robert Berls
Department of Education, Office of Planning, Budget and Evaluation

Delinda Cannon
Midland Technical College

Peter Capelli
University of Pennsylvania, Wharton School of Business

Magda Colberg
Office of Personnel Management

Elinor Miller Greenberg
EMG Associates

Richard Larson
Lehman College

Marcia Mentkowski
Alverno College, Psychology Department

Daniel Resnick
University of Pittsburgh, Learning Research and Development Center

Michael Scriven
Pacific Graduate School of Psychology

Mary Tenopyr
AT & T, Selection and Testing

Richard Venezky
University of Delaware

Joan Wills
Center for Workforce Development

Gary Woditsch
Private consultant and author

Steve Hunt, Assisting Recorder, OERI, Office of Research

Merrill Schwartz, University of Maryland, Recorder

WORKGROUP #4

Leader:

Andrew Kolstad
NCES, Education Assessment Division, Chief of Design and Analysis Branch

Participants:

Nancy Beck
Educational Testing Service

Wayne Camara
American Psychological Association

Allen Doolittle
American College Testing

Steve Dunbar
University of Iowa

T. Dary Erwin
James Madison University, Office of Assessment

Norman Frederiksen
Educational Testing Service

Ed Fuentes
National Education Goals Panel

Anthony Golden
Austin Peay State University

Joan Herman
University of California at Los Angeles, Center for Evaluation

Jeff Owings
NCES, Elementary/Secondary Education Statistics Division

Natalie Peterson
University of Pittsburgh, Learning Research and Development Center

Donald Rock
Educational Testing Service

Norman Webb
University of Wisconsin, Center for Education Research

Jerry West
NCES, Elementary/Secondary Education Statistics Division

Edward White
California State University, San Bernadino

Sheila Maramark, Assisting Recorder, OERI, Office of Research

Mary Carlson, American University, Recorder

POSITION PAPERS, BY AUTHOR, WITH REVIEWERS
(with ERIC numbers)

Position Papers/Reviews

1. Position Papers The position papers were designed to tap into the learning experiences of those involved in one or more aspects of assessing student learning. Fifteen authors were selected from eight different "Areas of Expertise" (noted below). The research and practical experiences of the author provided both conceptual, methodological, and practical insights for the creation of a process to assess the higher order thinking and communication skills of college graduates. The papers varied in focus and content, depending on the area considered. Each author was expected to include: (1) a brief introduction to the issue(s)/problem(s) under consideration; (2) a brief summary of the state of knowledge, including expected barriers and problems, from the perspective of the writer's expertise and experiences (as noted above this will vary depending upon the experiences of the author); (3) the author's specific suggestion(s) for developing and implementing a process at the federal level (this would include responses, in whole or part, to the set of issues and concerns noted earlier); and (4) the arguments for and against each suggestion with appropriate justifications. Each paper included a one-page abstract using the above format. (A writers meeting was held on August 30 to provide authors with a better perspective on the expectations of the paper content and presentation. Those unable to attend were provided with a report on the meeting.)

2. Peer Review Process: Each position paper was formally reviewed by three experienced researchers/practitioners. Reviewers provided written comments and related suggestions of their own. Readers also played an active role in the study design workshop.

Authors, were chosen for their interest/expertise in one of the eight designated areas listed below. They were not, however, constrained to addressing only those issues/concerns contained in the assigned areas.

1. Definition and Measurement Issues: Clarifying the definition and levels of proficiency for communications, problem solving, and critical thinking skills is a primary concern. How can the progress of students over their educational experience be assessed? Should assessment consider the value-added model of student learning or should students be expected to achieve a given skills level? Will this differ among students and if so on what basis? At what time in the educational process/life experience should students be assessed? Who should do the assessment and how often? Are there lessons to be learned from current tests and measurement activities and research?

2. State Experiences: Many states have or are planning to implement postsecondary assessment programs. The State of New Jersey has been notable for its postsecondary assessment activities. Florida, Virginia, and other states, have implemented state level assessment programs. What is the purpose of these programs? Who is to be included in the assessment and how are they selected? How is the information used in the states? Has it improved the quality of the educational programs? How can the experiences from these activities be used in the development and implementation of a national assessment process? How feasible is it to aggregate state level data for national assessment purposes?

3. Institutional Experiences: A selected group of institutions have been in the forefront of the modern assessment movement, assessing both program content and the general intellectual skills of graduating students. What can be learned from these experiences? What has been the focus? What are the skills tested, standards or levels of learning expected, and evaluation procedures being used? What use is made of the information? How can these experiences be used to assess the skills noted in this goal? How feasible is it to aggregate and utilize institutional based data for state or assessment purposes?

4. Relationship to Pre-Collegiate Testing: The skills to be assessed as noted under Goal Five for college students are similar to those noted under Goal Three for Grades Four, Eight, and Twelve. There are many approaches for the testing of these skills at the elementary secondary education levels. What lessons and suggestions can be derived from current testing and measurement activities at the pre-college level? What is tested? How broad are these tests? How valid and reliable are they? Can they be extended over the longer student experience? How are they administered? Do they lend themselves to a national assessment effort? Are they cost-effective? How fast can a program be implemented? What are the likely problems and/or limitations of each?

5. Testing Services Experiences: The American College Testing Service, Educational Testing Service, and others have testing and measurement activities focused on the assessment of basic learning skills at the postsecondary education level. It is important that the lessons learned from current testing activities be available. How does each approach the problem? What current testing instruments or procedures are relevant? What are the advantages and limitations of each? How would they be implemented? What are the costs and potential problems? What needs to be done?

6. Relationship to NAEP and the Adult Literacy Survey: Two existing NCES surveys focus, in part, on the issues to be studied. Each has a history of study experiences that can be a valuable source of information and suggestions for

the creation of an assessment model for the noted skill areas. These are the National Assessment of Educational Progress (NAEP) and as noted above, the National Adult Literacy Survey (NALS). What is the current or potential relationship of testing at these levels versus the postsecondary education level? Should there be continuity between the various levels in terms of definition, learning experiences, expectations, and testing procedures? Can either of the testing processes used in these studies be applied, in whole or in part, to this activity?

7. Job Skills Issues: American industry must improve its productive capability to keep up with the rest of the world. This means both enhancing capital equipment and related workforce skills. Job analysis is undertaken within the firm and by outside agencies. How does industry define and measure higher order thinking skills? How do they set standards? What lessons do they have for the development and implementation of national/state assessment of work and citizenship skills? How do they assess the skills of the college and non-college employee? What do they do with those who test below the standards?

8. Indirect Assessment Approaches: Some argue that there are alternative approaches to assessment, at least for the short run. For example, if one were able to identify the set of thinking and communication skills students achieved within a course(s), then the sum of the courses a student completes over the college attendance period would provide a rough indication of the skill levels achieved. Other proxy measures are said to be available. How valid and reliable are they? What are the pros and cons of each measure? What implementation and cost implications must be considered for indirect measures?

It should be noted that individual authors may not have responded to all of the general questions listed under the "Areas of Expertise" nor those summarized in the "Issues and Concerns" sections. Nonetheless, they provide, along with the reviewers comments, a comprehensive view of the issues and concerns that need to be considered collectively as an assessment process is developed. Since the papers and the reviews were distributed to all participants prior to the workshop, the proceedings effectively built upon the work of the authors and reviewers.

A listing of authors and reviewers follows. Copies of the papers may be obtained from:

ERIC Documentation Reproduction Service (EDRS)
Cincinnati Bell Information Systems (CBIS) Federal
7420 Fullerton Road, Suite 110

Trudy Banta, University of Tennessee at Knoxville: "Toward a Plan for Using National Assessment to Ensure Continuous Improvement of Higher Education." (ED 340 753)

Reviewed by:	Nancy Beck, Educational Testing Service Norman Frederiksen, Educational Testing Service Barbara Wright and Ted Marchese, AAHE Assessment Forum

Peter Capelli, University of Pennsylvania: "Assessing College Education: What Can be Learned from Practices in Industry." (ED 340 754)

Reviewed by:	Elinor M. Greenberg, EMG Associates Margaret A. Miller, Virginia State Council of Higher Education Mary L. Tenopyr, AT&T

Steven Dunbar, University of Iowa: "On the Development of a National Assessment of College Student Learning: Measurement Policy and Practice in Perspective." (ED 340 755)

Reviewed by:	John Chaffee, LaGuardia Community College Norman Frederiksen, Educational Testing Service Ronald Hambleton, University of Massachusetts

Peter Ewell and Dennis Jones, National Center for Higher Education Management Systems: "Actions Matter: The Case for Indirect Measures in Assessing Higher Education's Progress on the National Education Goals." (ED 340 756)

Reviewed by:	Robert Calfee, Stanford University Elinor M. Greenberg, EMG Associates Mary L. Tenopyr, AT&T

Charles S. Lenth, State Higher Education Executive Officers: "The Context and Policy Requisites of National Postsecondary Assessment." (ED 340 757)

Reviewed by: Robert Calfee, Stanford University
Richard Larson, Lehman College
Ronald Swanson, Texas Higher Ed Coordinating Board

Georgine Loacker, Alverno College: "Designing a National Assessment System: Alverno's Institutional Perspective." (ED 340 758)

Reviewed by: Elinor M. Greenber, EMG Associates
Margaret A. Miller, Virginia State Council of Higher Education
Mary L. Tenopyr, AT&T

Marcia Mentkowski, Alverno College: "Designing a NationalAssessment System: Assessing Abilities that Connect Education and Work." (ED 340 759)

Reviewed by: Richard Larson, Lehman College
Ted Marchese and Barbara Wright, AAHE Assessment Forum
Ronald Swanson, Texas Higher Education Coordinating Board

Ed Morante, College of the Desert: "General Intellectual Skills (GIS) Assessment in New Jersey." (ED 340 760)

Reviewed by: Richard Larson, Lehman College
Michael Scriven, Pacific Graduate School of Psychology
Ronald Swanson, Texas Higher Education Coordinating Board

Susan Nummedal, California State University at Long Beach: "Designing a Process to Assess Higher Order Thinking and Communication Skills in College Graduates: Issues of Concern." (ED 340 761)

Reviewed by: John Chaffee, LaGuardia Community College
Peter A. Facione, Santa Clara University
Ronald Hambleton, University of Massachusetts

Richard Paul and Gerald Nosich, Sonoma State University: "A Proposal for the National Assessment of Higher-Order Thinking at the Community College, College, and University Levels." (ED 340 762)

Reviewed by: Lorenz Boehm, Oakton Community College
Peter A. Facione, Santa Clara University
Ronald Hambleton, University of Massachusetts

James Ratcliff, Pennsylvania State University: "What Type of National Assessment Fits American Higher Education." (ED 340 763)

Reviewed by: Nancy Beck, Educational Testing Service
Joan Herman, UCLA
Ted Marchese and Barbara Wright, AAHE Assessment Forum

Daniel Resnick and Natalie Peterson, University of Pittsburgh: "Evaluating Progress Toward Goal Five: A Report to the National Center for Education Statistics." (ED 340 764)

Reviewed by: Nancy Beck, Educational Testing Service
Norman Frederiksen, Educational Testing Service
Joan Herman, UCLA

Donald Rock, Educational Testing Service: "Development of a Process to Assess Higher Order Thinking for College Graduates." (ED 340 765)

Reviewed by: Lorenz Boehm, Oakton Community College
Joan Herman, UCLA
Michael Scriven, Pacific Graduate School of Psychology

Richard Venezky, University of Delaware: "Assessing Higher Order Thinking and Communication Skills: Literacy." (ED 340 766)

Reviewed by: Robert Calfee, Stanford University
Margaret A. Miller, Virginia State Council of Higher Education
Michael Scriven, Pacific Graduate School of Psychology

Edward White, California State University at San Bernadino: "Assessing Higher Order Thinking and Communication Skills in College Graduates Through Writing." (ED 340 767)

Reviewed by: Lorenz Boehm, Oakton Community College
John Chaffee, LaGuardia Community College
Peter A. Facione, Santa Clara University

Peter A. Facione, Santa Clara University: "Critical Thinking: A Statement of Expert Consensus for the Purposes of Educational Assessment and Instruction" (ED 315 423)
Contributed Paper; No Reviews

Michael Scriven, Pacific Graduate School of Psychology: "Multiple-Rating Items." (ED 340 768)
Contributed Paper; No Reviews

THE STUDY DESIGN WORKSHOP

Compiled by Addison Greenwood

OPENING SESSION
Sunday night

SAL CORRALLO: Opening remarks and introduction of Emerson Elliott.

EMERSON ELLIOTT: Thank you very much, Sal. I'd like to welcome each and every one you, this evening, to the Assessment of Postsecondary Student Learning Study Design workshop. I appreciate your coming out on a Sunday evening, being willing to spend the next two days working on what is a really important issue that - I think - NCES has really not done enough about in the past. I passed another birthday anniversary this last week, and my wife found a really terrific quote that goes like this (We haven't been able to find the source, yet):

> "Age, cunning, plus treachery, will always win out over youth and skill."

So far as I can tell, that has nothing to do with this evening's conversation. Well, I'm really impressed, as I look around the room, at the many places that you all have come from, and the diversity of interests that are represented here. I'm especially impressed by the number of people from the West coast, who have come. I just returned from a meeting of the National Assessment Governing Board in San Diego, and I'm all screwed up on my time clock, so I appreciate you folks coming from the West coast.

San Diego, where I was, found itself on a list published in newspapers while I was out there of Best Cities/Worst Cities, from the point of view of their financial affairs. I don't know who put it out, but at any rate San Diego and Phoenix were tied for first place. And Dan Resnick and at least one other person here is from Pittsburgh. I appreciate your being able to come, from a city that was very close to the bottom, along with Philadelphia, Syracuse and Chicago. Whatever that means, maybe you like to come to meetings in beautiful Washington.

Well we clearly are here, I think, because of the National Education Goals. There could be lots of other reasons, and I'll talk about some other things, because I really think we're here for something that follows on a long term interest that has been growing over a very long period of time. And it's finally catching up to higher education. But for the moment, I think the place to start is the simple education goals that were adopted by the president and the governors of our states in February of 1990.

You've all heard these before, but I'm going to tick off the list of the half dozen areas, anyway, even though you have heard them before.

- That all students will reach school ready to learn.
- That we will have 90% of our students completing high school.
- That we will have a high level of student achievement: of students mastering challenging subject matter.
- That we will be first in the world in science and mathematics performance by the Year 2000.
- Adult literacy and lifelong learning is Goal Five.
- And Goal Six: all schools will be free from drugs and violence.

It's not just that the goals were adopted. The governors and the president committed themselves to being accountable for the goals, which they equated with preparing and sending to the American public, each year, a report that would accrue data on the progress that had been made by the United States toward that goal. And to carry that out, in July of 1990 a panel was established called the National Education Goals Panel, and Ed Fuentes is here from that panel. They were asked to determine, from all of the possible measures that might develop or are available or could become available, which are the ones most appropriate for following the project toward those goals? And if we don't have appropriate measures, to suggest then what form appropriate measures might take. The first of those reports has come out, on September 30 of this year, and if you haven't seen it, I think you'll find it an interesting piece of reading.

The National Center for Education Statistics has been very heavily involved in the preparation of that report. Sometimes in a visible way, sometimes in a not so visible way. It has provided staffing support for the White House, for the Goals Panel itself, for the various resource groups and task forces that are attached to the Goals Panel. We certainly have provided a great deal of data. Seventy percent of the data that was included in that report is from NCES work. We have reviewed all kinds of background materials and potential data sources, and have developed lists of new, potential

indicators. Internally, we have also tried to respond, as well as we can, to the needs of the Goals Panel, by doing special analyses, and responding to their requests, and including things in our data collection plan for the future.
Now the task at hand this evening, and for the next day and half follows up on Goal Five which says:

> By the Year 2000 every adult American will be literate, and will possess the knowledge and skills necessary to compete in a global economy, and exercise the rights and responsibilities of citizenship.

It doesn't say anything about postsecondary education. You have to read further because, in addition to the goals, there are some 21 objectives which in some cases provide additional detail on the goal(s), but in this particular case, stakes out something that appears to be fairly supplementary to what is included in the goal itself. This is objective Five, and it says:

> The proportion of college graduates who demonstrate an advanced ability to think critically, to communicate effectively, and to solve problems will substantially increase by the Year 2000.

Now in fact, while this has been included among the goals and objectives for higher education, higher education has not been singled out. This is virtually an extension into higher education of the same goal that is included for achievement at the elementary and secondary level for student achievement and citizenship for grades four, eight, and twelve.

There's a background paper that has been prepared by something called the Report of the Technical Planning Subgroup on Goal Five, that gives the recommendations that have been made to the Goals Panel itself while pursuing this area. I know that you have read this paper, but I think it's well to recall this at the beginning of this conference, because I think you have to struggle with these same kinds of things with which this task force has struggled. And this is what they describe as a system of sample-based outcomes indicators characteristics:

> It should profile graduates on scales for the full range of college-level achievement, not just basic or minimum levels of achievement.

It would not attempt to dictate a single standard of acceptable achievement. I do think one thing that pertains to higher education that does not pertain so much at the moment to elementary and secondary education is that people are not yet talking about achievement levels or standards. The point that I'm going to make here in a moment is that we're 27 years behind elementary and secondary--that's probably about when we'll get to standards in postsecondary education.

> Whatever assessments we have should use advanced assessment techniques that go beyond customary multiple-choice questions and include constructive responses, performance tasks, essays, and possibly even portfolios of actual work. The system shall provide information to guide the development of national higher education policy. It should not be used to compare, or rank, performance.

I think that's a very important distinction that the task force has laid out. And finally:

> Comparisons of scores across institutional types should be accompanied by information showing differences in backgrounds and abilities of students entering each type.

The National Center for Education Statistics has been involved in a series of assessment kinds of activities. Let me just tick these off, we are not newcomers to this area. The one everyone is most familiar with, obviously, is the National Assessment of Educational Progress (NAEP), the Condition of Education report, from NAEP data but a variety of other sources as well. There's the International Assessment of Educational Progress carried out by Educational Testing Service (ETS), the National Adult Literacy Survey (NALS), the National Household Education Survey, the 1987 Transcript Survey (also the 1982 Transcript Survey), High School and Beyond, the NALS '88 longitudinal study beginning with eighth graders, the Second International Science Study, the Second International Mathematics Study, and now we're preparing work on a study that actually was our proposal, the Third International Study of Mathematics and Science, to be conducted by IEA in 1994 and 1998. So we're not new to this area.

But I think conspicuously absent from that list is postsecondary education. The Center itself has identified some kind of postsecondary assessment as a gap in our data. We have discussed this on previous occasions with the American Council on Education network, with whom we meet three times a year. We are now exploring the possible use of GREs as one way to think about getting information on college level performance. But by and large we have not pursued this issue as aggressively as we might have. There is no question whatever that the National Education Goals Panel's interest has a lot to do with pushing things at them moment.

But I would like to indicate that the recent report "Education Counts," the report on education indicators from the special studies panel that was created to look at indicators, talked about getting information in higher education in each area where it's also appropriate to have it for elementary and secondary education. And one of the areas in elementary and secondary is student performance, and we should also do that for higher education. That report also talked about good schools, and said that we should have measures for good schools as well as measures of performance for students. It talked about links between the economy and education, and clearly that is of great importance at the higher education level. And it talked about equity as one of the major areas where we should have indicators, again for both elementary and secondary and for higher education. So that report is looking at some parallelism in our overall measures, for higher education as well as for elementary and secondary education.

In short, I would assert that, because of the Goals Panel, because of the growing interest, this national call, here, for a postsecondary student assessment is becoming more insistent. And I think that call is an extension of the move in elementary and secondary education that happened about 27 years ago, when Frank Keppell worked on the charter for the then Office of Education. It said that the Office of Education is supposed to report on the condition and progress of education and - he said - we never did that. And that cry came to Ralph Tyler and a few other folks to invent what became the National Assessment of Educational Progress.

Now I think that this move has to do with this general provision in our statute, that talks about reporting on the condition and progress of education: profiling it, describing it, monitoring it. And it is in a broad way related to accountability in American education. People want to see the results from funds that are spent. They want to see what difference institutions of higher education make. They don't just want to see that good students going in produce good students coming out. They want to see institutions making a difference. And they want assurance that the nation will create an educated and trained group of citizens that will make our nation prosperous.

I have not deluded myself about the difficulty of developing a postsecondary assessment. The interests are very diverse, and often conflicting. As you look at Addison's paper, and see in juxtaposition the things the different people who authored the fifteen papers have said, it's very clear that there are these kinds of conflicts.

The Center, somehow, as it contemplates this issue, needs to balance several things. One is, What analysts can tell us about what is possible and what is appropriate. And I think in the advice that many of you have given us already, we have some things there that don't exactly fit. And I think that will be a part of the conversation at this meeting. Second, we have to take into account what policymakers want. If you have talked to many policymakers, it is perfectly obvious that they have conflicting wants that sometimes don't fit together at all, but that is something that we have to deal with, since we are trying to respond to a variety of needs. Third, and I fell this very much from my conversations with the American Council of Education, we have to take into account what institutions will permit, what they will allow. And that, I think, is one of the most difficult parts of this whole assignment. And finally, we have to take into account what students will do. You might have an elegant design, but if the students refuse to participate, it will not be very useful for gathering statistical data.

Well this clearly is a formidable task. Formidable, actually, is the word - in the notes that Sal prepared it says "farmable," which I thought was really quite delightful. I was wondering if that meant that we would plant the seed, but the soil here is arid, and so it's really going to be very difficult to make this happen. But then I kept looking at it and thought he probably meant "formidable." Sal, I like that.

In preparation for the workshop, fifteen of you were asked to prepare position papers. I'm told by Ron that this is a stack about so high, is that right? - at any rate, it's a large stack. Each author was asked to consider four areas of concern that really are the organizing principles of this conference, and they are: What should be asked? Who should be assessed? What standards should be applied? And how should the assessment be accomplished? Fifteen more of you reviewed those papers, and identified issues, which are included in Addison's paper, that really are what this conference is all about.

We hope to use this meeting as a way to start a national dialogue on a postsecondary assessment of critical thinking at the collegiate level. That is something that a statistical agency must do. The Goals Panel has this task for itself, and I think is looking at it very much from a policy perspective. But I think the additional requirement for a statistical agency - if we are going to gather data - somehow we have to be pushing very much harder on a lot of issues that policy people can gloss over. What is it we are going to measure? How are we actually going to do it? What kind of questions should we be asking? And will anybody participate in this party if we do ask them?

What we need from you is to tell us where we should start. That is, What should be the next steps? On the basis of the papers that have been prepared to date, I think these kinds planning exercises involve a lot of steps. It's not simply taking fifteen papers and saying, well, that's the plan. It's obvious from the fifteen papers and from Addison's summary paper that there are a lot of things that still need to be looked at. I think, following this meeting, I was talking to Sal about this during dinner, the sort of thing that inevitably will happen is that there will be a report back from each of your groups, and from the group as a whole, and that needs to be available for a wider public conversation.

But then I think that the National Center for Education Statistics must do something that statistical agencies always have to do. And that is to figure out how to take all of this advice and turn it into a design for something. And then that design itself needs to be made available for further public comment. I do see this as something that will take several iterations, I don't know all of the iterations that were undertaken for the development of the National Assessment of Educational Progress between roughly 1964 and 1969, but there were a lot. And there were a lot of compromises made, and people had to be very clear, on the one hand, about what it is they were trying to do, what this neat new thing should be; while many other people made it clear what they would permit to happen and what they would not. And for the last five years we've been busily changing all of that, which was apparent in this meeting I just returned from, in California.

Well we need a plan of attack, and it does require planting seeds, and it is a formidable task, and I wish you all well during this conference, and I hope to join a great deal of it as you proceed. Thank you very much.

SAL CORRALLO: Well, he's laid out, the task for you, to plant those seeds. And hopefully someday this will grow into the kind of project where we'll all be proud of having been here. (After a briefing on the activities planned for the rest of the workshop the group adjourned for the evening.)

SMALL GROUP MEETING REPORTS

The small work groups remained distinct from on another throughout a long, 12 hour working day, with the exception of informal exchanges in the hallways during breaks and during mealtimes. During the working group's deliberations, group leaders--where it seemed useful--tried to structure their group's discussions along the lines of four sets of questions posed in the introduction to this working paper. The four group leaders varied with respect to how closely they followed the outline of questions, and how much importance they gave to each topic. The work group reports have been reordered to match the outline of questions. As with any such reordering of comments, possibly rich ideas are prone to fall through the cracks. Indeed, the topics themselves came more into focus as thinkers probed and poked at their edges and overlap. The distinct contributions of the four work groups are preserved as they were arrived at - separately. Interestingly enough each of the four groups produced clearly distinguishable reports, reflecting no doubt, the personalities as well as the backgrounds, of the group's members. The final results summarized here, and necessarily this condensation, tends toward common views and consensus. Where repetition occurs from one summary to another, it marks a nonetheless independent discussion among different groups of participants. The source reports from the recorders remain at NCES--do the original notes and tapes--with all of their authentic detail, debate, and diversity.

At the end of the day, each group leader undertook--in collaboration with group members--preparation of a summary of their groups deliberations for presentation the following morning to the assembled participants. A transcript of those reports is presented first.

Summaries of the group deliberations follow the leader's reports. They were developed from reports prepared by the two recorders. One recorder summarized salient points on large sheets of paper that were posted around the as an ongoing record of the discussion. A second recorder, while taking notes, used a tape recorder to capture the major points in greater detail. Subsequently, these several records were combined to produce the reports presented below. Quotations are informal, paraphrase is assumed. The goal is not attribution, but rather to evoke the spirit of the discussions. Most of the phrases and all of the ideas presented in these reports come directly from the discussants.

GROUP LEADERS REPORTS
Tuesday Morning

SAL CORRALLO: This morning we start with twenty minute reports from the work group leaders. I'm told they will allow minority reports from other group members, but within the twenty minute framework. No more. After a short break we'll open the discussion to the floor. As you can see we are also recording this session.

A few words about the closing exercise. Each of you has in your packet two forms. One asks about exemplary practices. If you know of any that you would like to identify and nominate, we will pass that information on to others in OERI who are working on the process. We're looking for "what works" kinds of things. FIPSE funded projects are good examples of things that we would like to publicize in that document. We have a FIPSE rep here with a big smile.

On the second form there are the four items we've asked you to respond to. Last night I said you could mail in your comments. Yes, you can, but I would like something from you today. Even if it's rough. You can follow that up with other information, anything at all, that you may want to send us. Any ideas, we're open. Don't - at this point - send us proposals for funding, we aren't in that business. Send those to Cliff Edelman. We're looking for ideas, and obviously some research that will come out of this. So, we're in the open ears stage: take advantage of it.

Finally, at the end of the morning, Gary will be answering the question, "Where do we go from here?". Emerson will also offer remarks, but if you know Emerson, you know they will come at the proper time. However the proper time may be during the discussions or in closing. Emerson will be the final voice today. So with that, we'd like to get started, let's take Group Four first, catch him off guard, and go in reverse order.

ANDREW KOLSTED (Group Four): Before we addressed the four questions, we though a little bit about the contributions of assessment to improving instruction, which we thought of as something this whole enterprise ought to be directed at doing. To improve teaching and learning in the secondary area. We spent some time on that, but then, without coming to any particular conclusions, we then moved on to the four questions.

The first one on Domain. We started on the issue of wondering if the three things identified - CT, PS, and CM - are these the most appropriate qualities that we should be assessing? They're not completely distinct. There's some overlap among them, and we know that CT has some components, including meta-cognitive approaches, and the ability to recognize alternative perspectives from that of your own.

In terms of the question, "Is there another concept or subject that we ought to be measuring besides those three?" we didn't come up with any significant element that had been left out. No one thought another subject needed to be added to those three.

We did discuss, "Are these abilities (CT, PS, CM) generic, or are they discipline-specific, and how tied are they to specific disciplines?" We agreed that a generic assessment is necessary and can be done. Is feasible. But we did note that often background knowledge - and a very high level of background knowledge - is needed to be able to demonstrate CT. So we though the assessment should consider tailoring subject-specific parts, which might be different for students in different majors.

So there might be a core design, with some questions for everybody, and replaceable modules for people in different fields. How broad those different fields might be, whether there would be five different one for things like natural sciences, social sciences and humanities, or more narrowly defined modules - maybe 30 or 40, one for each department - we didn't resolve that. We thought that was something for later on, but the basic idea was a core and then replaceable modules. That covers our discussion of the Domain of the measurement.

Then we got into what subjects should be tested. We had five different populations that we considered. The four-year college graduate. These are essential,. but we discussed whether we should be testing prior to graduation or after graduation; there are some advantages to either. Testing people after graduation would assure that they had graduated. We wouldn't be competing with the demands of school, finishing up all their course work sometimes people are the most busy.

However, this is offset by the fact that it would be more expensive to get at them. If we test people before graduation, they are easier to locate, and we can do testing at central sites that would be more convenient for administrative purposes. How we proposed to resolve this would be to try or test both approaches and see which turned out better. It might require developing two different designs in order to test the two methods. That's the first population.

The second one was to consider whether people with A.A. and A.S. degrees were also college graduates, and we decided that yes, the two-year and community college graduates ought to be tested as well. Then we thought, What about a comparable group in the four-year colleges. A majority were in favor of including second-year students in four-year institutions. We also talked about entering students as a population to be tested, from the point of view of looking at gains, but we didn't reach any agreement on this, and thus make no recommendations.

Another population that could be tested, again for purposes of having a comparison group or showing what college does, and that's the non-student, same-age cohort, who are not in college at all. But we considered it, and basically rejected it. We though that we didn't want to test them. That covers the five populations.

What we then talked about under this heading of Who to Test, we talked about "What are the motivations of people to participate?" For faculty, we expect some resistance, and thought that feedback on the findings of the study might help to overcome that resistance. For students, we thought research would be needed on what might be effective. The presumption seemed to be that we have a lot to overcome in encouraging students to participate.

We came up with five different kinds of things that can motivate students. One is high stakes. That is, require a test in order to graduate, or to get a grade, or something like that. It works, but I don't think we can get that into our study. The second was financial incentives. But people also suggested three others.

One was giving them feedback on their performance directly. And while for some people this was a motivator, for people at the low end of the scale sometimes that's not a motivation. The other two were public recognition of the fact that they had participated in the study. That might have some value. And the last one was just a general climate that students are expected to participate, and often respond to the expectations of being good students. So that covers the second question: Who should be tested and why they would want to be tested.

The third issue we discussed was standards. We came up with two qualitatively different kinds of standards. The first one: proficiency level descriptions. We thought that we needed to be able to describe what the scale means, at different levels of the scale. What performing at a certain level indicates in terms of the kinds of things that people can be expected to do when they score at that level.

So this means behaviorally-anchored to items of a certain kind that we might have three to five levels described. That these descriptions would depend on a cognitive model, which, since it's an empirical model would have the possibility of rejection of data. So that's all in terms of descriptions of what people can do.

Then the second kind, the qualitatively different kind was evaluative standards for what's not good enough. We thought that these standards come from a policy or political decision. While there is an empirical part that involves the gathering of data from experts, in terms of what their criteria are, gathering judges by itself is insufficient to get reliable judgements of what's good enough or what isn't good enough. And we thought that we ought to defer making such judgements until much later in the course of this project, and let a political body do it, rather than us.

And then, the third point on the Standards issue was that we thought more research is needed on how to combine expert judgements and evaluations and setting standards, because the methodology for developing reliable judgements of standards - including all the judgements of experts - is not that well developed yet, and needs considerable work.

This brings me to the last of our four issues, What instruments should be used for assessment? We though that the assessment would profitably combine multiple approaches. We thought that the traditional multiple-choice items would play a minor role, but not a major role. Part of it might consist of that, but there were a number of people who thought them sort of incompatible with measuring CT. One person offered some experience with multiple-rating items, which are different from multiple choice, but are still machine-scorable, to get around some of the problems of giving away some of the answers in the coils of the alternative responses.

The best approach we thought was an extended, free response approach. There was a good deal of sentiment for going substantially beyond the traditional one or two hour assessment time. Some people thought we should go all day. Others thought perhaps we could send people home with a problem, and have them come back later, having thought about it, sort of a take-home exam idea. There was a minority of one who thought that portfolios - taking advantage of extant work - is a feasible and possible thing to use.

In addition to all of these direct measures, we also discussed - but didn't reach agreement on - indirect measures. By indirect measures we mean such things as self-reports on opportunity to learn, motivation, and participation in curricular activities and extracurricular activities. We did discuss, as I mentioned yesterday, transcripts as a sort of indirect measure of these things, but decided that there was so little content available in terms of what constitutes the course that we didn't think it would be a good idea: we rejected that as a measure.

In any case, whatever measures we develop, the group thought that the measures need to be validated against the theoretical descriptions of what it is we're trying to measure. Because many of the existing measures of CT, on their face, don't look that close to real life demands or don't seem that valid.

I tried to cover things that we all agreed to, but Sal has reminded me that there may be minority reports, so if there are people from my groups who would like to volunteer additions, or remind me of changes, or point out things where I didn't summarize, this is an opportunity to do that. (No volunteers.) Well, I think that because about ten or twelve of the people in the group came up to help me make those corrections last night, I'm not surprised.

JEAN GRIFFITH (Group Three). As I talked to my fellow facilitators, I realized that the discussion in the four different groups did in fact go differently. There were different comments raised, and different concerns, and it kinds of reminds me - when I think about what NCES is going to have to do to try to put together all of these different opinions. It reminds me of the veterinarian and the taxidermist who went into business together and put up a sign that said "Either way you get your dog back." So, we're going to have to put all of this together, and figure out what the product is going to be at the end.

In Group Three on most issues we had a wide range of perspectives, reflecting discussion about a wide range of options for a national post-secondary assessment system. We had a great deal of concern expressed that the system should be dynamic. That there should be aspects of assessment, of associated information, and of related indicators. We talked about the four areas that Andrew just used, and that we all constructed our discussion around, and I will focus my comments around those four areas, and I'll add in a couple of other areas that we discussed in some depth. What I'm going to try to do is to identify those areas where our group was pretty much in agreement among ourselves, and then I'll talk about the areas where our debate did not actually lead to resolution.

We talked about why we would be doing a post-secondary assessment. We felt that addressing that question is really going to be critical for NCES. Our ability to present a convincing argument about why it is important to do a post-secondary assessment, why we are developing such a system, is going to be very important in making it acceptable to the community that will be affected by it.

We felt that the post-secondary assessment system should go beyond accountability. It should be providing feedback to the post-secondary education system, both in a general and a specific way. Feedback to policymakers and faculty. We felt very strongly that the system need be grounded in educational purposes, that it should be used to improve instruction and learning.

We thought that we needed to develop a multi-dimensional system to address these twin issues of accountability and instructional needs. And finally, as I mentioned yesterday I believe, the system should reflect the contributions of education to the economy, but also as well to the society and to the politic.

In areas of disagreement on why we are doing the postsecondary assessment, there was disagreement as to whether or not the system should feed back, directly, and in a very precise way, to institutions, to faculty, and to students. And there was some debate about how assessment actually gets linked to instruction and learning.

We talked about the domain of the study, what it is we should be doing. I think there was general agreement that a profile of general skills could be assessed, that cuts across disciplines, and should be embodied in any college graduate. We felt there was a need to assess active learning that reflects thinking processes, self assessment that interacts with feedback, and other behaviors related to learning. We also felt that we should be assessing the transferability of skills among disciplines. And also between education and the world of work.

There was a lot more debate about the specifics of what we should be doing. People weren't certain how deeply we should get into the area of general skills versus assessing in various disciplines, how to balance the assessment program in those two areas was not resolved. There were specific issues about how a learner learns, and how a teacher teaches, and how to link those issues within an assessment system.

We talked about indirect measures of assessment, at some length, and by this we meant indirect indicators of assessment. We were not able to create a set that we thought would be useful, and which would not be dangerous were they to be misused. It seemed that every idea we came up with engendered debate about whether or not that might not be an advantageous indirect indicator. So, what we decided was that - if such a system were to be developed - it would require considerably more research. And we would need to create indicators that had little anticipated potential for misuse, and also that we could not be able to anticipate serious unintended consequences of the indicators. We were very concerned about consequences of indicators that would cause people to alter policies to change the indicator, but not to change the underlying system of improving education.

When we talked about Who should be assessed, there was agreement that all people in some age range should be assessed, so that we should be assessing graduates of two-year institutions, four-year institutions, proprietary schools, and also the non-collegiate population: people who enter the military, the workforce, and people who work at home. There was discussion, though, about how to accomplish exactly that, whether you would use an institutional survey, in addition to an age cohort of the population, to attain the kinds of information you would need.

We talked about assessing with a longitudinal versus repeated cross-sectional measures. There was considerable agreement that we needed to have measures of people at an age before they would be entering post-secondary education, as well as following post-secondary education, in order to understand the contribution of post-secondary education and other life experiences to improving the kinds of skills that we felt should be assessed.

There was debate about the sampling issue: what level of the population do you strive to attain representation for? Do you try to get a sample that's representative at the national level, the state level, institutional level, for faculty, for individual students? When we talked about that, the discussion was couched in two separate areas.

One was sampling representativeness, and the other was the groups to whom you provide feedback. We separated them because we felt you need to think about how the sample represents. If the sample is not representative at the faculty level or the institutional level, but you have state level information, you can use that state level information to provide feedback to the individual institutions, faculty, and students who participated in the program to help them understand what the program is about, to motivate them, and to provide positive feedback for their own instructional purposes and educational development.

We also talked about the longitudinal aspects of the survey. About assessing lifelong learning. And we acknowledged that implies a much longer-term assessment program, and the acquisition of quite different kinds of skills. I think that although we discussed it, the general sense was that this project could not serve to make assessments of lifelong learning.

We talked about our instrumentation, and I think we had some of the same discussions that Group Four did, in this area. We were in agreement that we need background information; we have to be able to describe the characteristics of who is being assessed. We thought that we would need information on individuals, on teaching characteristics, on learning, and on institutions. However, we had a lot of debate about that, about exactly what kinds of issues should be included in the body of information from the survey for research on these issues.

We decided that there has been a fairly rich body of research developed on these issues, and that NCES needs to consult that quite carefully as it develops the background instrumentation for the survey.

We also talked about the importance of the linkages between education and the world of work: that is very important to our group. Trying to understand the relationships between what we're doing in post-secondary assessment and what the Dept. of Labor is doing with SCANS and other projects, is something that we clearly have to take into consideration as we are developing our own national assessment.

When we talked about assessment processes, there was no sentiment in our group that I was able to discern in favor of straight multiple-choice testing. We also talked about the multiple rating items as a viable and very attractive strategy, that represents an important advance over straight multiple choice. We also felt there needed to be some kind of a writing sample in the project.

We talked at some length about performance assessment, and felt that it is an extremely important method to achieve validity and to measure the types of abilities that we are trying to target in this assessment. There was some concern about the ability of performance assessments to adequately tap the content domains, and we felt that it was important to strive for generalized ability in this assessment. And so there was some concern about using a complete performance assessment to achieve that goal.

Finally we talked about Standards. And we considered how to anchor standards in an assessment, and considered three different approaches.

The first approach was a norm reference system, which nobody thought was the most constructive approach, in this case. We talked about anchoring in everyday performance, and we felt that what that would lead to would be averaging across the demands placed on students graduating from college. Our concern here was that we would be focusing on what is currently and actually being demanded, rather than what can be, what we can produce through our post-secondary education system.

So we ended up thinking that we should pursue the third approach to anchoring, which would be to use human expertise to identify expertise in certain fields, in the ares that we were assessing. Where you would identify the abilities of outstanding people on a scale, and this brings me right into the next point about standards. That is, that we felt very strongly that they should be developmental. That there shouldn't be single cut-offs. That we should talk about performance over a range, on a scale, and yet that still needs to reflect expert opinion about what the highest levels are that we should all be striving for on those skills.

If we adopt standards, we felt that they need to be variable. By that we meant two things. They should be variable over time, the standards need to be responsive to changes in the system. We have high expectations that this assessment system would feed back into the educational process, would improve the educational process, and consequently, standards would be able to be raised on some kind of a periodic basis.

We also felt that standards should be variable among populations, so that, in terms of our sampling - we are sampling students graduating from four-year, from two-year institutions, people in proprietary schools, people who have not attended post-secondary institutions - any standards developed should reflect the different skills, talents, and experiences that people from those different groups would bring to the assessment.

That summarizes what I have to say. does anybody in the group have anything to add to that, or to modify what I have said?

MARCIA MENTOWSKI: I'll just make one comment. I think we talked about how standards, if they were descriptions of the abilities we were trying to measure could be a mode of communicating to the general public, and to faculty and to students, what it is that we were about. It could be a prime motivator for participation. Some kind of picture of the ability could be communicated through standards if they were not quantitative but rather qualitative descriptions of abilities.

JEAN: Exactly. Thank you.

RON HALL (Group Two): Sal, in changing the order, you destroyed my introductory remarks. I thought there would be a nice point/counterpoint with Group One, based on the discussion we had with Gary yesterday. Actually, Groups Three and Four covered a lot of the territory we did, but I think there are some important differences.

First, I would like to characterize the group, because I think it's important and I think others feel it's important to the way that we addressed the task, and some of the agreements and disagreements that emerged. In looking around the table in Group Two, I noticed that these were all people, at the state and institutional level, who have been doing assessments. Conceptualizing, designing, building consensus, and implementing and evaluating these kinds of assessments. And I think that they brought a particular perspective.

Consequently--a second point I want to make--is that there was a tension that lasted all day. From the opening speaker until 9:45 last night, we focused on the underlying tension inherent in identifying the propose of assessment. We labelled the dichotomy--as did some of our other colleagues--improvement versus accountability.

Some of us took positions on the basis of the intent of the goal itself. Is it an accountability goal, or is it intended to be a vehicle for improving student learning? Are we to use the results of assessment to change what we do in institutions, or does the goal's importance lie in answering the questions of the public. Such as the return on the investment in higher education, or the overall quality of undergraduate teaching.

In the end, and I will return to this briefly as a point of agreement, there was general consensus that -despite the overwhelming difficulty - the two contexts, improvement and accountability, are not mutually exclusive. There is optimism that both can be achieved, despite the operational problems. So, unlike Group Four, this notion, this dilemma, characterized our entire day's discussion. We did cover all four question areas, though not necessarily in order.

Let me turn to what I think is Group Two's major contribution. In the afternoon, we disciplined ourselves to focus on the four questions, and ended up with a proposed process for development of a national assessment. It has these features. It begins with an institution-based assessment. Grants would be provided to a limited number of institutions that volunteered to be part of the experiment. Twenty-four or so was mentioned, but that number is certainly not sacrosanct.

It would be initiated quickly. We have fears that other kinds of pressures may grow very rapidly to push our education institutions into an assessment that won't work. It would have two basic tracks, and these two tracks would operate somewhat simultaneously. There would be an outcomes assessment track, and a research and development track. Basically, here's how it would work.

Institutions that volunteer would select from a limited set of existing or new, serially-designed, devices or systems, and try them out. Say a set of three to six (existing) plus some new devices or systems. By the way, I'm using the term devices or systems, because this group wanted to get away from the notion of a single test. One or more coalitions of institutions would be encouraged to develop - from their assessment experiences - their own model. The outcomes assessment track would be cross-referenced by the research track, perhaps by short-term longitudinal studies, or pre-post-assessments.

How do you get from this proposed assessment development process into a national assessment? Well, after a two or four year period, the experiment would be rigidly evaluated - and there were some references to some of the deficiencies in the Kellogg (sic) experience) - and that exercise should lead us to some options, which I think are listed over here.

- o Briefly those would be that we might be able to pick one or two of the assessment systems or devices that stand out as being successful in a variety of institutions.

- o The second possibility is that we might pick a set of such devices or systems - and this option obviously causes us to have to look at the question of commonality across the set - and go the route of combining the best elements of each of these devices or systems.

- o And lastly we might discover, through the process, that there are sufficient commonalities across the whole lot that we could use them all. This of course implies a cluster approach to assessment, when you move it to the national level.

Well, having outlined this process for developing a national assessment, let me turn to some areas of agreement along the lines of the four questions that were posed for deliberation. These are also listed on one of the charts, but not necessarily in the order I will give them.

The group achieved consensus around the notion that we should assess all three areas: effective CM, PS and CT. And that we support multiple indicators, including perhaps indirect indicators. The group believes that the characteristics or parameters of the domain can be identified as follows. I'm going to have to read these from the chart.

In-built process for flexibility
Relevance to the real world
Skills integrated with content
Skills incorporated into performance
Skills incorporated into practice
Recognition of all levels
Common standards versus diversity (each with its liabilities)
Must be sufficiently complex to require multiple performances

Now having essentially laid out these characteristics, the group felt very strongly that practitioners should be the ones to lay out the operational specifications for the domains. Along with that there was strong feeling that we need to develop a deeper understanding among faculty in institutions about these skill areas, and how to teach them.

The group achieved consensus that we should assess students while they are still in school. And we, like the other groups, would like to see that extended to two-year schools as well as four-year. This agreement was sort of arrived at negatively, by rejecting as part of the main assessment certain other concepts. Such as international comparisons, assessing a non-collegiate population, and a post-graduate sample.

With regard to standards and instrumentation, the group feels there is much valuable experience already out there, upon which we can build. Hence, our notion in the development experiment that institutions should include existing assessment devices in the experiment.

The group is definitely opposed to a multiple choice approach. The primary reason for this is the superior validity of a performance-based design. The group believes that the importance of authentically assessing the kinds of higher order abilities developed in college is worth some sacrifice in traditional concepts of reliability. The kinds of skills addressed are manifest in complex performances, and should not be separated from those performances.

Finally, a point of agreement, as I noted earlier. The group feels very strongly that both improvement and accountability can be approached together.

My final word is that there are a number of areas requiring further research. I will just read these out, as they were noted in our discussion:

What is the nature of the top-down versus bottom-up approach to assessment design and implementation?

What is the strategy for change? This came up late in our conversation, when we suddenly realized that in all the deliberation about the goal and all the talking that's been done, we haven't rally specified that strategy for change. So, at both the institutional and the national level, what is the strategy for change?

A lot of discussion about how to motivate students to show up and do their best, and I'm sure the other groups addressed that too. But also we would be interested in further research on motivating faculty to get involved in this sort of thing.

Where, and at what level, do CT skills get developed? And how do colleges contribute to this? How do non-college students compare?

Those were the basic points, and I'd like to invite my colleagues from Group Two to clarify, add, or give some minority views. Peg?

MARGARET MILLER: I wouldn't say we actually rejected the idea of an international and non-collegiate populations. WE thought of that as a project that shouldn't be rejected but rather postponed, as ancillary, later in the process. Second thing is, I'd like to say a little bit more about the relationship between the assessment system we're proposing and existing assessments. Because, it seems to me, one of the features of this proposal is that it doesn't wipe out what's already been done in assessment in the various states. And particularly those other assessment processes may be more powerful for understanding just how - in any particular set of circumstances - change will occur in the teaching and learning process. So it would be very important I think, throughout this process, to encourage existing assessment processes to continue.

RON HALL: Any others? Trudy.

TRUDY BANTA: It's certainly not a matter of disagreeing with what Ron has said, and said very well. But it's a matter of emphasis, and I think that from what has been said so far the emphasis on what we are doing here, and on what may be done from here on is on assessment, when in fact the statement of the goal is in terms of improvement. And it seems to me that what we talked about was a strategy for changing what is happening in the teaching/learning arena that would improve student learning. And if the real goal here is to increase students' abilities in each of these areas, then it seems to me that what we ought to be focusing on is strategies for improvement. And using assessment as a means of checking our progress in that regard. And so the idea that we would start with some development grants is a very exciting one to me. We would check our progress using the kinds of assessment instruments that we have been using in the last five to ten years in secondary institutions, and so assessment would definitely be a part of it. But the focus would be on strategies for improvement. One other thing. I don't think we want to throw out the multiple choice option entirely. I think we might find that is a very important part of any assessment system. We certainly wouldn't want to turn it all over to that option though.

RON HALL: Any others from Group Two? The debate continues. . .

GARY PHILLIPS (Group One): One of the things I tried to figure out, as I mentioned yesterday, is that we thought Group One would be highly contentious and full of trouble makers. I think I found out what happened. There was apparently a pre-meeting of the theoreticians the night before, and I think they got drunk and had some sort of Eureka! experience, and realized that basically they agreed on these things.

To start out with, I think I disagree with something that Jean said. I don't think it was a taxidermist and a veterinarian. I think instead it was a critical thinking expert and a proctologist that got together, and the sign said "Either way we improve your dispositions." Those theoretical critical thinkers will especially appreciate that.

In Group One, I think we took a broad but not a very deep approach to the various issues, which I think is appropriate to this stage of the project. And we sort of methodically went through all four of the issues. So let me go through our thinking on these issues. The first issue had to do with the domain of interest. And the first question we asked was "Can this be done?" and "Should this be done?" and the answer was "Yes" to both.

The next question had to do with "What are the desired outcomes, or the skills, that we should measure in post-secondary education?" and the general feeling is that we're on the mark: it is CT, PS and CM that we should be looking at. Then we asked ourselves the question "Is there a set of core CT skills on which a national consensus can be obtained?" and the answer is "Yes," there is a core set of skills, but it would not include all CT, PS and CM skills, but there is an important core, and a national consensus on that can be obtained.

And "Are these generic skills, or are they connected to certain specific disciplines?" And the answer is that they are generic skills, and there are measurable outcomes of these skills. So this made us feel sort of good, that we can proceed to the next ???

There was also agreement that we need to not only talk about these skills, but we also need to expand the assessment so that it's more comprehensive, and it covers the dispositions of CT. These are the habits that we use in everyday life, such as thinking independently, intellectual perseverance, being clear, questioning our assumptions, appreciating evidence, this sort of thing. And so there was a feeling that we need to go beyond the simple skills and abilities and also cover these dispositions.

We also felt that there should be an extensive set of background questions. Demographic information about students and instructors, transcript information, information about the instructional processes and other activities that go on in the institutions.

"What are we trying to measure?"

If we could do a quality assessment, then a longitudinal type study is always better than a cross-sectional type study, because it gets at the value added by the college experience. This allows us to answer the question "DO colleges make a difference?" And it helps eliminate the problem of self-selection, of students going into colleges. But if we cannot do a quality longitudinal study, then we should do a quality cross-sectional study. It was simply a matter of which can you do best, given the funds. Everybody agreed that - if at all possible - this ought to be a longitudinal study, with, at the minimum, a pre-major going in, and a post-major going out.

What is the level of aggregation? There was a unanimous view that we should not be comparing students, institutions, or states. Instead there should be a national assessment, with of course breakdowns by types of institutions and that sort of thing. But institutions and states and students should not be identified.

We talked about the unit of analysis, and we all agreed that the unit of analysis should be the student. That's not the unit of reporting, but it is the unit of analysis. What we're interested in is student CT, not institutional or some other unit of analysis.

Should we include outside populations, outside of the colleges, the military and the general population? The answer is "Yes" we should do that. Should there be an international component? Our feeling was that 's a good idea, but we ought to defer that until we get this thing off the ground, and work on that later.

We then turned to the third issue, which had to do with standards. There was agreement that "Yes, we should set proficiency standards" although we did not discuss how. But again, we should only do this if it can be done correctly, and in some satisfactory way. And also that the standards should be set not on global composites, but using as fine a grained assessment as possible. For example, CT has various components, and if possible, there ought to be standards set on those components.

One other idea that was discussed and agreed to. In addition to - or maybe in lie of - setting standards on some kind of scale that would be empirically determined, one other approach, particularly if we have a dominantly performance assessment, is that the standards could be set within the scoring rubrics themselves, rather than on the scale. This is what is done currently in many state departments for direct writing assessment. You have the standard right in the scoring rubric. So you actually have it before you even give the test.

On instrumentation, we discussed for some time about whether a single test can be used to monitor as well as to improve educational progress, and I think everyone recognized the problems. But I think the consensus was that we should proceed anyway, and sort of worry about this as we go along, and not let this stop us. In terms of the role of multiple choice versus performance-type assessments, there was general agreement that - as much as possible - we want this to be a performance assessment that can simulate the real world situations that CT actually occurs in.

And a range of performance assessments appeared to be appropriate: written responses, since writing is CT. Oral responses, portfolios, at best samples. A need to have items that require elaboration. Explore the use of computers and other technology and that sort of thing. The general rule is: As much as possible we should make this a performance assessment, but not rule out multiple choice testing, right up front. And also explore multiple ratings that were mentioned earlier, and other possible technologies using multiple choice testing.

Can extant instruments be used, for this assessment? The general feeling is that of course we need to look at what's been done before, but there probably isn't anything off the shelf that could just be borrowed and used for this purpose. And there was a lot of talk about making sure that the people who work on his project thoroughly understand the successes and failures of previous projects and other instruments.

Should matrix sampling be used? Yes, it was agreed that matrix sampling should be used. This gives us broad coverage. And also we agreed that we need to proceed in such a way that we continue to meet the joint technical standards of the American Psychological Association, the AERA, and the National Council on Measurement of Education, and also the Code of Fair Testing Practice. It was acknowledged that these standards are not particularly rigorous, or that they're not as detailed as we might need, and they also don't cover performance assessment very well, but still we need to be in line with these standards. And as a matter of fact, NCES has committed itself to follow these standards, so that's another reason.

We did not want to discuss what model could be used, what sorts of models were available. But one model that was mentioned was the National Assessment of Educational Progress (NAEP). And one of the things that the group liked about NAEP is that, as it's currently configured, it is in three different components. One part of NAEP is a set of surveys that gives an instrument the same way each assessment, so that you can monitor trends over time. This is the concept that "If you want to measure change, you can't change the measure." So we keep giving the same instrument every time, which helps us monitor trends.

There's whole other set of surveys and instruments in NAEP, which have to do with innovative approaches to assessment. New concepts about what should be measured. And that's a whole different set of surveys, so there are two big pieces. That is where you're trying out new things, while you're still measuring trends in the old way.

And possibly most important for this project, is the third component, the research and development, and statistical research studies that are going on in NAEP, for example, oral reading and meta-cognitive strategies in reading, and this sort of thing. And so these are small scale research projects, the results from which don't really have implications until perhaps five or ten years down the road in the assessment.

And the feeling is that we need in our project to have a similar model, where we would have something like those three approaches. Particularly, up front, we should be funding a considerable number of R&D studies, for example ethnographic studies, studies that have to do with the obstacles to CT, teachers and students, multiple ratings, and other approaches to multiple choice tests, and meta-cognitive studies.

Are there any questions, disagreements, or elaborations? Richard?

RICHARD PAUL: I just wanted to emphasize a point that didn't come out clearly in what you said. The CT researchers are very concerned, based on our experience in California, to be involved not simply at the theoretical construct level, nor at the item instrument level, but also on the interface between the actual design for testing, and the reporting back stage. Because I think that at each of these stages, if there isn't feedback from researchers solidly grounded in the substantive concept of CT, there's room for a slip to be made. And since we've experienced such slips before, we're especially concerned to be on the record of calling attention to that possible difficulty.

GARY PHILLIPS: Richard, I'm glad you mentioned that. We've had that same issue come up with the national assessment. Where the people who are involved in the initial development of the assessment are not involved at the stages where items are written, scoring is done, reporting. And so they're unhappy with what comes out at the end: it wasn't what they had envisioned. We're fixing that in the national assessment, and I'm sure that as this project proceeds, we'll do the same thing here. A very good point. John?

JOHN DALY: One thing I wanted to bring up. I come from the communication profession, and the program so far is primarily CT, and I guess one of my hopes is that we don't consider this whole program as a CT exercise. Clearly there are massive overlinks between the three, but there are two people from Communications here, and as far as I know, no one from the area of PS. And while those are somewhat related to each other, they are also somewhat independent research areas or disciplines, and independent instrumentations. And it seems to me that one thing we need to spend a lot more time on is the communication aspects of this, whether it be speaking, listening or writing, as well as the PS things. Most of the presentations today--while they're very, very good--use CT as the key term. And (if you listen to the tape) most people are talking about CT consistently, yet the goal talks about the three: there are commas between them and an "and," suggesting three separate areas that would need to be examined. So I think in probably any conclusions that we reach, there needs to be a stronger focus, both on the CM aspects as well as the PS aspects, in addition to the wonderful work being done in CT, which I think is very, very important. But that's one of my concerns all the way through. We're always using the term CT, and then we throw in CM sometimes, and PS seems to have simply disappeared from the discussion. I'm not sure why that is, but we don't see people bringing up that area very much.

GARY PHILLIPS: That's a good point. I think in future activities on this project, we'll be more attentive to having a more representative sample of people in the three different areas.

SAL CORRALLO: Is there a problem solving group in this country? I assumed if you're a critical thinker, you know how to solve problems, so maybe that was an incorrect assumption. Please, in your submissions, put the names of people who you think can represent that school of thought. We are very interested in identifying people from areas that we don't have fully represented here.

A SYNTHESIS OF WORK GROUP REPORTS

As noted earlier, each work group was unique in the time and emphasis placed on the topics covered. The following synthesis was developed around the responses to the main concerns underlying the set of questions posed to the authors, noted in the Introduction. These included the meaning or rationale for the study, the domain or skills to be assessed, the subjects or who is to be assessed, the standards or levels of achievement to be defined, and the instrumentation or alternative approaches to assessment. Although the intent is not to summarize, the restructuring does provide the reader with a tool to review the deliberations within the small work groups. Work group reports are presented in reverse order to match the group leader reports to the body.

GROUP FOUR

Group Leader: Andrew Kolstad

Recorders: Mary S. Carlson, Sheila Maramark

The Meaning

While bringing a somewhat less eclectic background to their deliberations than did Group Three, Group Four nonetheless produced less consensus than they did insight into how two particular subgroups may view NACSL from different perspectives, that of critical thinking experts and measurement people. It may have been a rendition of what a joint meeting of Groups One and Two would have produced, but with a substantial footnote. Ed White, whose paper advocated the utility and power of portfolios in assessment, continued to carry that banner in the context of a group whose members showed some support, but no little skepticism, about portfolios.

These group dynamics are mentioned as they seemed to offer a possible insight into how some of the central issues surrounding development of NACSL might get resolved. Problem solving is not only one of the 5.5 target abilities, it was one of the central activities of the workshop. How did Group Four (or all groups, for that matter) solve the problem of deliberating the issues raised by the NACSL? To oversimplify, the "measurement people" (Participants in this work group included a larger number of people with hands on experience and training in the process of measurement) were not inclined to elaborate in detail on the particular solutions (instruments) proffered. This may be a reflection of their professional experience, where they create and administer instruments only after all of the variables have been identified and specified. Such specification was obviously not a given, and so they joined into the debate over the other questions with zeal, understanding (perhaps better than others) that when the actual instrument comes to be constructed, many questions now unresolved will have been answered.

Alternatively, the "critical thinking people" (Group One) inhabit a particular discourse community, one where the questions of definition, purpose, population, method, and utilization of results seem to have been resolved, or at least addressed, to the satisfaction of people in that community. Whether or how portions of that approach will transfer into another context not only remains to be seen, but could also be seen as problematic in the same way that what Alverno does is seen as problematic.

The Domain

Predictably, Group Four displayed a fairly rich and complex view on the 5.5 skills. As named, they represent a starting point, but in practice and in the literature, CT, PS, and CM are not wholly distinct from one another, nor can they be properly embraced without including metacognitive perspectives. But this way of describing the domain is acceptable, to these experts, and does not omit anything significant. In particular, CT is often viewed as a tactic rather than a testable skill, and may be invoked by subjects in dealing with almost any testable task, that is solving any problem. If the CT people want to further distinguish, they may use the term "effective thinking" in an attempt to provide a tighter definition and more accountable framework.

The big question: Should we test for discipline-specific or for generic abilities? The answer, neither can be excluded. A better question is to focus on which majors (and how tightly they are to be defined) lend themselves to which types of subject-sensitive examination, assuming the goal is to assess CT and PS.

The Subjects

Let's correct the phrase "college graduate" right now, suggested many. The impracticality of trying to assess people outside of and after they have departed the college context is patent. Four-year people will be tested, but so should two-year graduates, in the same way, if with a different test. A majority of Group Four would also test the four-year people at the two-year level as well, for baseline and for comparison purposes with other populations. This might provide a better baseline than testing entering students, but a number of considerations--mostly practical ones--about a longitudinal study were contentious. It was generally agreed the non-college age cohort was not likely to be captured any time soon, though "if we want to demonstrate the impact of college attendance, testing non-college attendees could be very useful . . . so little is done with these kinds of comparisons . . . this could be a real opportunity."

What about motivation? Clearly it is a problem, and while many of the group had considerable experience, insight, and recommendations to make, as a whole they recommended that the literature be consulted and that structured research be inaugurated. Feedback is important: to students to provide them clear expectations, structured motivation, and recognition for progress; to institutions, where faculty resistance is a real problem to anticipate. A model which received much discussion and favorable reviews was that of James Madison University in Virginia, where one day in the spring is dedicated to assessment. Among other benefits, faculty have a vital role in this process, and have generally learned not only to assess but also to teach in a more holistic fashion. Creating a positive environment for assessment is not an overnight proposition, accomplished by fiat. Such a process often "takes three to five years to become part of the intellectual environment of a department."

Standards

Group Three saw this as two more or less separable questions: the one technical, the other political ("The real issue here is a political one, who establishes what's good enough.") To help mitigate that problem and misuse of the NACSL, use the model of "proficiency level descriptions." Examples of this are needed, but they should be flexible, and anchored in behavior. Preferably a minimum of three or better five levels of scoring to guard against high stakes and political misuse. Empirically, there is insufficient research on this extant, and while judges should be a part of the process in order to bring stake holders into the process, real experts are necessary to conduct the sort of valid research necessary, first for methods to set standards, and then to provide guidelines for the standards themselves.

The problem, of course, is to provide a system that can be made consistent and replicable across the diverse populations and institution-types involved. But the process should begin by establishing the best of what is known about assessment. Then the more pragmatic elements of how to create a national instrument and how to insulate it insofar as possible from political abuse and manipulation can be addressed. The NAEP experience should not be lost, especially with respect to trying to balance the differing uses to which the data can be put: to establish trends or provide individual feedback.

Instrumentation

To repeat, Group Four included a number of testing professionals. Thus the discussion was rich with detail, if not clear-cut consensus and recommendations. A diversity of methods is a given, with straight multiple choice questions having only a minor role, multiple rating items a more major one. Notwithstanding the inherent difficulties, extended free response items that might extend well beyond a couple of hours were thought potentially valuable. The portfolio idea, strongly advocated, won little support, as people felt they were uneven, unwieldy, and perhaps not relevant to NACSL as presently conceived.

The idea that at the present stage we should be looking for a collection of indicators rather than a single test was not without its appeal. Several felt that it would be worth while to better quantify how useful extant materials might actually be as at least partial indicators. Conversely, the view was expressed that the less intrusive the (surrogate) measure, the less incisive and revealing it was likely to be. There was a call to refocus on the essence of problem solving, to realize that a problem is often more than just an assignment to be completed, but rather something in the real world that is by definition ill-structured. The consensus believed an instrument could and should be constructed, but that a flexible and multiple approach to this task was crucial.

GROUP THREE

Group Leader: Jeanne Griffith

Recorders: Merrill P. Schwartz, Steve Hunt

The Meaning

In sharp contrast to Group Two, Group Three was composed of educational thinkers from a number of different backgrounds. This diversity reflected itself in a healthy and often skeptical debate about their very charge. Many of them foresaw prodigious conceptual and operational difficulties with NACSL. They did not consider closely too many of the positions and arguments offered in the authors' papers, though five authors were in the group. Consensus was rare; concern about symbolic and operational difficulties was constant. A number of group members had very strong positions on many of the key questions, and expressed these forcefully against the occasional skepticism from such a disparate group of colleagues. This summary will be most useful by trying to include some of the stronger positions, both pro and con, rather than to impose a balance and consensus that was never achieved. A number of key questions for OERI and NCES were identified, but not necessarily answered.

On the central conundrum of the day, they believed serving the twin masters of improvement and accountability will be difficult, but that NACSL should be grounded in improving the 5.5 skills. Nonetheless, they debated with some heat whether direct feedback to students, institutions and/or states was the way to accomplish this. They called on NCES to put forth a strong rationale once begun, believing that the philosophy articulated at the outset will inform NACSL throughout its conceptualization and implementation. They felt strongly that restricting the standards by which the project will ultimately be judged to improving the economy was too limiting, and that a stronger link needed to be made between the goal of improving CT, PS, and CM and the ultimate benefit to society.

The Domain

Unable to agree on whether the instrument(s) should test for general skills or rely on subject content, the group ranged over the underlying strengths and weaknesses of both extremes, and of combination-type tests. Of particular concern was how the tested skills would transfer, and they seemed to suggest that even the literature isn't unambiguous, and raises the considerable specter of active vs. passive learning. An important question is to see whether passive learning is fundamentally inconsistent with CT, and if so to discover how to weed it out of NACSL.

A second point of concern was that performance results on the test might be misused, and therefor that somehow we need to create a structure as resonant as possible with the underlying goals. In this view, assessment results are not synonymous with life trajectories. Therefor the question of what indicators to develop and/or look for was addressed more widely than merely to inform us about the actual content of questions on an exam. Others mentioned include how employers structure opportunities; what happens to workers on the job; how communities use libraries; the impact of community characteristics on the means and the opportunity to learn; data on opportunity to learn in college (e.g. who requires theses? which majors correlate to success?); certain labor market indicators; and indicators within the educational infrastructure, such as instructional practices, class size, ACT/SAT scores, and grade point average in high school.

The Subjects

The test population should be as broad and inclusive as possible: two-year and four year college students, public and private institutions, the military, and the non-college age cohort. This philosophy also suggests that lifelong learning is an inextricable part of the process, and some felt the study should be longitudinal for years to come, tracking students through their careers.

If a sample of students is selected, be very careful about the process used, and establish a panel of experts to assure validity. As to motivation to perform, the group kept referring back to their earlier discussion about indicators, hoping for a neutral and non-threatening collection of indicators, but unable to suggest its outlines at this point.

Standards

It was around the question of how to arrive at and to implement standards that Group Three seemed to coalesce. Their most forceful and positive recommendations were manifest in a view that standards have an important role in NACSL, namely, to encourage the aspirations of students. Necessarily, these standards will be developmental, and will reflect different institutional types. Nonetheless, the best possible human performance should be elicited and set as the top of a scale, performance on which will continually be related to other variables. The process by which these standards and variables are discovered and codified is crucial: it must be broad, public, inclusive, and should foster wherever possible partnerships between the stakeholders. The very process of arriving at, refining, and continually revising standards many saw as a new communication tool, one that could not only foster learning and improve instruction, but actually cement the success of NACSL itself.

Instrumentation

Group Three preferred to look a bit more broadly at "How" NACSL might come about. They did, however agree that the traditional multiple choice test was in and of itself inadequate. Consistent with their developmental goals for the assessment, they wanted many options to be considered, and as many stake holders as feasible included in the process: multiple rating items, writing samples, in-basket exercises, group problem-solving tasks, interviews in-person or over the telephone, portfolios, performance based open-ended questions and employer fault-tree analyses.

GROUP TWO

Group Leader: Ron Hall

Recorders: Christine Carr, Jeff Gilmore

The Meaning

Group Two was composed of a great many "testers," that is, people whose combined experience in test development and administration embraced a great deal of recent American history in the field. This background seemed to provide them a different starting point than the other groups. While their consideration of some of the more esoteric and philosophical issues raised was less extended than that of other groups, they moved aggressively to envision a real-world model on which to proceed.

Illustrative of this pragmatic approach, little time was spent anguishing over the compromises and trade-offs involved in trying to balance improvement of instruction and accountability. They stated unequivocally that both are essential, and must be incorporated. Thus, the question wasn't whether, but how to accomplish this. The answer, at least the slogan: Improvement is the goal of NACSL, accountability is the means by which this shall be achieved. Avoid a structure that can be analyzed--or criticized--as an either/or Hobson's choice. They understand the public's questions must be answered. And they know that the opportunity presented by the national goals and NACSL is precious: if they don't give the stakeholders what they are asking for and will be satisfied with, the effort could come a cropper. But they emphasize that this procedure will not be a quick-fix, rather the re-orientation of the process. An enlightenment for both the institutional infrastructure and the public about--given what NACSL reveals about the state of the target skills--how to begin to think differently about systemic, ongoing, constructive change.

But why not provide those answers in a context and framework where the target skills will necessarily be enhanced? Because of their close familiarity with most of the major state and federal precursors to NACSL, they were ready and anxious to propose a working plan. This strategy turned out to be very fertile, in the sense that very specific (and yet flexible) steps were sketched out. Because they have experienced many of the scars of battle in other efforts, their warnings were usually accompanied with suggestions on how to make it work. One major point was to realize that only the

U.S. and Canada among developed countries place the burden of higher education on the student. Thus, while international comparisons are premature as to the actual outcomes of the assessment, the point should be emphasized that a major federal role is the modern standard among our global competitors.

The Process

Let's begin right away, they said, transforming the test development process into an evaluation of a set of possibilities that are tested in the trenches. A number of separate efforts should go forth during this period, many of them embraced under what might be termed a "research track." Since they themselves embrace a great deal of assessment experience, they would like to see a process established whereby the "right" questions could be asked by way of evaluating the existing knowledge. For example, as begun in Charles Lenth's paper, what does the state experience tell us about what works, and possibly about why what doesn't work fails? Similarly, a way of evaluating the contributions and models embodied in certain institutions, such as Alverno, in the context of developing a real-world national model.

One strong inference already in from both of these "databases" is the bias towards the local, institution-based, faculty-guided assessment. Ergo, an important early question for the research track: How can we aggregate and establish a national reference system for individually tailored assessments? During the day, the idea developed that a sample of institutions--perhaps one to two dozen--would voluntarily participate in the NACSL development by implementing early versions and ideas as suggested by advisors brought together by NCES. Thus, suggestions and possible features of NACSL could be constantly fed into the embryonic process, and feedback just as consistently and quickly modify ongoing experiments.

The Domain

Sensitive to the creation of a system that won't work on the very local level, Group Two wants domains to be derived from the "bottom-up," that is, with a strong institutional flavor. To begin with, suggest to the institutions that volunteer for this research experiment that our overall philosophy is for an incisive, diagnostic probe, rather than a broad aggregation of skills and abilities across a great many students. That kind of broad charge dictated, then provide only the broadest parameters for domains, albeit emphasizing what is known and learned about how CT, PS, and CM seem to be involved, and can be approached. Then look to the institution and the faculty to either select from among known instruments and approaches--or to modify and devise their own. The watchword from above is flexibility: institutions are encouraged to find creative ways to uncover a system that will work, for them; and from the federal scientists should come an encouraging, non-polemical, non-judgmental collector of good ideas.

The Subjects

The insistence on a longitudinal study found in other groups was here relegated to the research track. On an experimental basis, absolutely, try to test students at the one, two, and four-year levels. But more important to this group was how to identify, and then to motivate, a sample of students that would be truly representative, and from whom valid generalizations might be drawn. How to define and select this cross-section was thought to be problematic, but crucial: another important task for the research track. Representativeness was important not only within, but also between institutions. For the experiment, one should ask for volunteer institutions, but establish a system with enough incentives so that many colleges will apply. Then be careful to select an array of institutions for the experiment that will not be as ecumenical as possible, with respect to structure and approaches to education.

The group's experience prevented them from reducing the complex issue of motivation to platitudes or bromides. First, there is the question of motivation from the institution to the students. Experience suggests the best large structure is to administer the test to intact classes. How individuals approach the experience most feel is strongly influenced by the messages delivered by faculty and the school. Thus, a structure needs to be found where the school has incentives that individual faculty can also perceive and embrace. The most obvious, though complex approach, is to establish the assessment as an integral part of a given course's successful completion. This raises the specter of fairly elaborate and complex feedback relationships. But students need to be able to see a direct relationship between their efforts on the test and their approach to their course(s). This also relates to the earlier point about bringing faculty into the heart of the judgment and development processes.

Another issue that blends both motivation and whom to test relates to how large the sample must be in order to fairly reflect both course choices and the demands certain curricula make on students' CT, PS and CM abilities. Careful analysis of curricula needs to be done, hopefully in a way that will produce a formula that can then be generalized throughout the country. A best guess now is that 10 to 15 percent of students might be selected to represent about 80% of the curriculum. Ultimately, it would be useful to compare this group to the non-college and the international populations, but for the present stage(s), that seems unrealistic.

Standards

This group manifested a very complicated attitude towards standards, perhaps because their experience has shown them that such a large variety of standards exist and get applied in different contexts. The ultimate question they saw as: Can we derive a common standard? For now, pending the results of the research track experiment, the answer seems to be "only in the most generic, flexible way." They envisioned standards as sensitizers, ways of making the domain-specific instrument choices better fit the population(s) being tested. Across different institution types and across different domains of testing--and especially across time--standards need to be continually refined and re-evaluated. Such an approach also dovetails with their emphasis on giving the institution and faculty more access to refining the instrument for their own purposes.

They earmarked some particularly thorny problems. Grammar, both of test questions and responses, is an issue that cuts against people for whom so-called standard english may be viewed as a "second language." By continually emphasizing the NACSL as a national indicator, rather than a "national test," standards can be used not to discriminate against--but rather between--special populations and those with certain skills. They favor the way NAEP was originally conceptualized, in that scores were not intended to show deficiencies to an absolute standard, but rather where in the range of desired outcomes particular people or groups fell. Though difficult, developing an instrument whose domains are fully integrated into performance, and then making adjustments for content-specific factors, seems to be the best way to avoid the misuse of standards. If such a process is continually fueled and periodically revised by the input and approval of practitioners and higher levels of stakeholders, the chances for misuse are further lessened.

Instrumentation

As part of their experiment, Group Two wants to see the coalescence of a consortium of participating universities that would act through a representative governing body to develop NACSL. The research track would focus on identifying and evaluating extant instruments in use around the country. The flexibility principle in this context suggests the use of what was called a "toolbox" approach. A core set of possible instruments would be a part of the basic choice, which institutions would have some freedom to select and to modify as they found indigenous ways to refine and improve them. The central body would continually monitor the use and effectiveness of these piecemeal approaches, trying to narrow in on one or a small group of most effective tools. The end product would be a recommendation, which itself may continue to include some element of selection for participating institutions once the real NACSL is underway.

Clearly this approach emphasizes flexibility and vests considerable choice and decision-making in the participants, whether states or institutions. Also, investment is made in improving extant instruments rather than designing a novel approach which would take much longer, with no real reason to believe there is some new eureka approach not already in part realized in extant thinking. The process needs to be controlled to the extent that the ultimate outcome is an instrument (or small array of choices) that can be used nation-wide and provide data that meets the needs of accountability, as they become more clear.

Some guesses as to how it might work. A dozen to two dozen participating universities, given two years to experiment with various approaches and revisions. Be thinking about validity as the process is unfolding, and make certain modifications when necessary to further identify promising approaches. Present stakeholders need to be consulted as well as brought into the structure. Central body is not delivering expert opinion and dicta, rather is supporting a joint effort at practical research.

Ed Morante, whose paper on the New Jersey experience provided a real basis for establishing certain basics of the discussion, suggested the outline of a process: Define your ultimate goal. Define operationally what CT,CM, and PS, are to be. Take a look at various current procedures that are available that come closest to meeting the operational definitions. Select top three (devices). Put out a notice to institutions: we will fund you if you try out these devices ($25,000 plus expenses?) Encourage faculty at some institutions to explore what kind of techniques are appropriate to do these things, before the final instrumentation is selected, so we get that part going. Collect data and begin the process of validating which one or more instruments seem to be reliable, valid for assessing the goal originally defined. Replicate that. Take information we're getting from the institutions, and begin to clarify the kind of skills students have at the level we decide we want. Then show change over time.

GROUP ONE

Group Leader: Gary Phillips

Recorders: Monika Springer-Schnell, Pat Dabbs

The Meaning

Why NACSL? Throughout the day, people returned to the political and philosophical reasons underlying their consideration of a national assessment. "We need to look at--and be able to prove--whether or not college makes a difference." Defining the concept of American higher education, pragmatically, translates into "**naming** the desired outcomes of postsecondary education." If we're going to try to say "what American postsecondary education is providing students, then there must be some measure of change over time."

And, in what was to be a common theme and a recurrent focus across groups, tension seems unavoidable between the strictly pragmatic view ("Determining the level of student knowledge and skills is the primary purpose of assessment." "The assessment process must begin. A national assessment process cannot be done behind closed doors. There is political pressure to get this underway. The train is going to leave the station.") and the more idealistic charge ("An important secondary purpose of assessment is to help make improvements in American higher education.").

Closely allied to this dichotomy is a concern, rising to the level of apprehension in some, for how NACSL might be used politically. One point later elaborated in the Tuesday contribution, was "What happens to the ball once it leaves our hands? What happens to the scholarship that went into deriving the instrument? Is the process going to be politicized, diluted, and/or restructured to the point that it is something we regret having participated in?" The group seemed to agree this was a danger, and that the best antidote was to be sure the process "needs to leave room for self-critique and approval, rather than being carved in stone and inflexible. There needs to be room for multiple approaches, tracks, or models." As one put it: "We cannot look at assessment in terms of a "one size fits all" instrument. Ultimately, however, policy makers want national information from assessment."

Be careful, warned the group, about a number of potential pitfalls. "We must be wary of the "value-added" principle. If Harvard students have the highest scores, it does not necessarily indicate that Harvard results would serve as the best pre- and post-measures for assessment." "There is a need to address questions of ethnicity, gender, age, socioeconomic status, and region of the country" in developing NACSL. "It is important to distinguish between what is maximum performance versus average/typical performance of people. There is a need to identify what we are measuring here." "It is important to get the right normative perspective. If we do not take something that is not just right or perfect to conduct a national assessment, we risk that we could get something worse. In any assessment, it is not possible to assess everything."

In sum, awareness of the political realities--in part shaped by experiences in other situations--made the group wary of developing and then endorsing a NACSL that could be misused in predictable, political ways; or one that might be too flawed by unrealistic, albeit idealistic, hopes of informing and improving the postsecondary system and its participants. The NACSL should not undertake to evaluate "each college/each person in the United States. This is both unfair and is not doable. Let us hope that [the train now leaving the station will evaluate] the state of education and the "readiness" of our young adults as aggregate information. This is how they are prepared for these kinds of skills in these kinds of institutions. We do not want to see that a student does not graduate based on a score on a national assessment instrument. We also do not want to see an institution having its charter revoked based on the scores of its graduates."

The Domain

What should be tested? As Richard Paul and others with a rich and substantive view of CT were in this group, it wasn't surprising that one recommendation was "that we look at gathering more support for the proposed list of skills for assessment," as detailed in the synthesis and original paper(s). While this and all groups considered the content domain from various viewpoints, many of them generic, a recurrent theme was the need for further research and testing of the content domain, as the NACSL development process proceeds.

A major question often revisited was "Where are the links of assessment to the real world?" Social science has one answer: "The links of assessment to the real world need to be made through content validity. We need to look at what we do and then analyze what we do in terms of critical thinking concepts. We need to make the link of conceptualizations of critical thinking and its application to real people." "We need to look at how skills are manifested in the workplace and in citizenship. It is real world tasks versus abstract ideas of critical thinking." "We must get employers to use the results/outcomes of assessment."

The skills, ultimately must be articulable, with outcomes that are measurable. They must thus be identifiable, and demonstrably relevant--"They must count... They must make a difference." As to the relevant CM "skills, certain skills have long term predictability in terms of outcome measures. They are core skills on which national consensus could be obtained and for which predictability would exist."

The group addressed the question of just which outcome measures might be considered. It was generally believed that no such consensus could be reached on either of two mutually exclusive approaches, the one involving generic abilities, the other subject-specific. "Both play a role. They are articulated in different ways with somewhat different emphases." It was thought important to "study the domain specificity issue. There is a need to look at "fields of activity," or the extent to which there are differences/similarities in looking at critical thinking within particular subject areas." Metacognition was a recurrent theme of group one: "Metacognitive skills become important when looking at the transfer of thinking skills across subject areas." How does one begin to specify this? "We need to analyze the course syllabi. What is there to facilitate the kinds of skills that we are talking about?" Also it was stressed that some methods work better in the intra-disciplinary context than across disciplines, and vice-versa. But there was a consensus that grade point average (GPA) was not track with or reveal CT, and that students themselves were well aware of this.

This group's experience in analyzing skills from the substantive perspective was manifest in their concern about quality of thinking, which brings to the forefront self-assessment and self-reporting, and crosses over to the later issue of standards. CT is a process, and manifests a certain quality. Students "ideally know how to assess their own critical thinking. To be interested in critical thinking is to be interested in intellectual standards. The two really cannot be separated." One should "treat measuring of standards as assessing quality of thinking." To do so necessarily entails thinking about thinking--metacognition--and developing a CT approach to PS; that is, approaching a problem in a way calculated to yield the best answer. Students must be looking at themselves, demonstrating awareness of evidence and analysis strategies, showing the ability to develop defensible analogies, and showing that they know--beyond simple matching of multiple choices--why an answer is right or wrong.

Outcome measures entail a number of hidden phenomena, and group one was very concerned that these be addressed. One such is institutional. Different models may be needed for community colleges and conservatories than for four-year universities. Developing the CT and PS skills is in part a function of time, and a two-year process is necessarily to be distinguished from a longer one. Other measurable distinctions might involve age and gender.

Another buried but crucial element of developing a measure involves what may be classified as cultural influences and attitudes. In particular, how such phenomena may block CT, or qualify its measurement. A credo: CT cannot be measured out of context (not to be confused with CT measured in a specific subject or discipline). This context should become part of the database for subsequent analysis, seeking information on attitudes, self perceptions, personality characteristics, family background, language used in the home, and other ethnic background factors. Sociological factors are also relevant: family socialization patterns, television and other media in the home, socioeconomic status, job experience, and career aspirations.

The environment at the school cannot be ignored. Attitudes of faculty are crucial. "Contextual information for assessment from students, parents, and teachers is needed. Teachers, for example, would be asked specific questions about their training." Studies need to be done "on attitudes of faculty that relate to critical thinking. For example, there is sometimes a faculty attitude of critical thinking being reduced to the mental hardware of IQ. Look at the extent to which faculty think their subjects foster critical thinking automatically. There is also a need to look at the extent to which students feel they can rely on cramming to pass and to look at student attitudes toward the intellectual as facts and opinions. There is a need to analyze the concept of reasoning and reasoned discourse. Documentation of attitudes is needed in both the faculty and the student populations to see the blockages to critical thinking." These attitudes are manifest, and may be measurable, buy looking at peer socialization patterns outside the classroom, at the general environment in which student life is experienced: is it conducive to CT?

The Subjects

There is a strong preference for establishing a longitudinal study, rather than a cross-sectional one. The group appreciated the complexities and the cost involved. One possibility--not simple but perhaps economical--was to select a sample from a currently funded, ongoing longitudinal study, to link some extant data with NACSL. (Another suggestion was to take a longer-range view by testing now, and testing again the same subjects eight years hence.) A longitudinal study is the only way to verify that the CT skills elicited were improved in college. Moreover, the baseline established in such a system would provide a strong impetus for improving instruction, the other concern consistently raised in discussion. As the variation in students' college experiences is related to their assessed performance, natural reforms in the system should follow. The view was expressed "that if we are doing a reasonably well designed assessment, we will probably find that we are doing worse in this area of higher order thinking and communication than we think we are. This may get the financial backing for working on improvements."

The quality of the assessment should be foremost, and other practical adjustments made to this principle. Matrix sampling, and probing in various ways the results of each student assessed was favored. A consensus was reached, in part driven by the need to mitigate the appearance of a "high-stakes" test: the test should have a longitudinal design, with a national focus adjusted with the value-added approach. The unit of analysis should be the student, and institutions should be considered by type, rather than individually. Comparisons within states, between institutions and between individual students are to be structurally discouraged. Reporting must serve the federal master, of course, but be designed to obviate predictable political problems that have been seen in other situations.

Standards

It was agreed that "defining the *x* can be a real problem in higher education," and especially that the kinds of CT, PS and CM skills being sought are inherently hard to quantify. The group warned that proficiency scales tend to become canonized. Such a predictable phenomenon needs to be anticipated, lest a monster be created that violates the very principles of CT. If a given score comes to be publicized and linked with a specific level of developmental deficiency, some of the basic premises of CT are violated. Given this group's professional investment in CT research, they voted strongly in favor of approaching the question of standards with a formal CT study that would involve experts from the field. The goal would be to develop paradigms that could be used to guide the test designers.

This is sensible, given the NCES exercise, because the CT movement has developed certain working principles, foremost among them the concept of collaborative learning: using the test to provide feedback to students who then go on to develop a profile. Over time, quantifying elements of this profile can have a positive influence. Thus by producing a composite score, you begin to mitigate against political misuse and a counter-productive reporting system. An aspect of such a system is to involve instructors in the judging process. The group noted this was essential in any event, to respond to the charge to assess all three of the 5.5 target skills. This approach suggests a kind of "multi-dimensional space" in which students performances across a number of skills and dimensions are related. "There should be a strong emphasis on functional relationships" between and among such separately tested skills.

Some practical considerations arose: Given that a simple numerical result is inadequate and misleading, how will the data be aggregated and scored? How will it be collected, whom will it be reported to, and for what ostensible purpose? These larger questions cannot be divorced from first identifying and then implementing the necessary standards into the assessment. (E.g., if the result of a certain set of scores is to direct state money towards remedial education, and that money is available, then standards developed in the instrument will in effect implement this larger goal.)

Two other issues arose, reflecting a concern for standards: the question of citizenship, and the feeling that essay writing was essential to a proper NACSL. With the former, it was believed that "there are certain public issues that one should be able to discuss at a certain level" of awareness and sophistication. As to essays, notwithstanding the practical problems, it was felt they provide a rich source of data from which a variety of profiles could be developed. This leads to the practical problem of designing an instrument.

Instrumentation

Again was echoed the overriding political awareness of how the test might be (mis)used, the so-called "Lake Woebegone effect," by teaching to the test. High stakes thinking and thus countermeasures are unavoidable. The group noted an irony, however. If the tendencies for high stakes are ultimately irresistible and indefensible, then why not try to "design a test in which you get a teacher to teach in just the way you would like him or her to be teaching." This idea is attractive, but very complex, and suggests that instructors become much more actively involved in the implementation and the judging of results than before.

Even if a portion of the test is to be subject-specific, it is important to raise the target to CT skills, not (e.g.) math or science per se. The test must include a multiplicity of indicators: multiple choice sections can focus on CT micro-skills; multiple ratings yield more esoteric and coordinated skills; essays and extended responses are crucial to getting at these higher skills; performance is necessary--writing and speaking must be demonstrated, preferably in simulations that are authentic and mimic real world situations; portfolios can help in this search for practical results. Again, matrix sampling can provide an economical way to develop high quality measures. The group warned against trying to re-invent the wheel. "There are well-established, effective testing modalities now." Depending on the time frame and long range planning, "There are also innovative testing strategies under development, presently at a more seminal stage now that show promise for long-range, future use." But extant instruments should be reviewed for item types, and a hybrid instrument developed with elements of most of the above.

CLOSING COMMENTS

Participants were invited to offer individual comments in an open session. Although rebuttals were accepted, few challenged one another, rather preferring to offer additional advice or reinforcing suggestions made earlier. No attempt has been made to summarize these comments since they build upon ideas and suggestions made within the work groups. They will be factored into the development process as the subject is considered.

My name is Bob Ennis. I'm from the Illinois Critical Thinking Project at the University of Illinois, and have been interested in this problem for a long time. In 1958, my doctoral dissertation was entitled "The Development of a Critical Thinking Test." I've been working on the problem ever since, and am delighted to see the great interest that has developed in this area.

The burden of my comment today is to try to resolve this basic issue of the purpose of this operation. Is the purpose assessment, or is it improvement of instruction, as the issue is put. And I want to suggest that it can be both, but to do so in a way such that some of the problems that would develop if the same administration were used for both purposes might be avoided.

The problems with the use of the same administration of a single instrument for both purposes: If you use the instrument (or the device, or whatever we may call it) for the improvement of instruction and learning directly, then many more students would have to be tested. I should say that I am assuming that our basic goal is to get a very high quality set of devices or instruments to use, in order to make the judgement that there either has or has not been substantial improvement in these three areas. That's my basic assumption.

Now if we use the same administration which is used for making this judgement for the direct improvement of instruction, then we'll need to give it to many, many more students than we would otherwise. This will increase the cost astronomically--this I can speak to with great assurance, from my experience--and thus be strongly motivated to compromise [reduce] the quality. There are all kinds of things that will happen if we try to give this to many, many students and make it a high stakes kind of instrument.

One way that the quality would be reduced is to reduce the amount of personal interaction between an evaluator and a student. Another is that matrix sampling might become impossible, or at least very difficult to do. It's hard to run an experiment and to give feedback to a particular student if that student is only one small part of a matrix, for example. Now if we use this administration for improving instruction, then we will have to derive information that is by student, by institution, and by state. Such a structure will make it a high stakes operation, with the newspapers reporting scores, and institutions comparing each other, and thus all of the attendant attempts to try to teach for the test, or perhaps to have certain students not take the test. You know all of the devices people use when we use high stakes instruments.

Now I think there are four ways, at least, that we can get an instrument or set of devices that is primarily a monitoring kind of instrument. I also want to distinguish between monitoring and accountability. Monitoring is really what the goal calls for. It doesn't call for us attributing responsibility to the states or to institutions, which is what accountability would do. If we just know what the level is, then there's much less chance of high stakes. So, instead of just having two choices, accountability or educational improvement, I think we really have three choices: monitoring, accountability, and instructional improvement.

What I would like to urge is that we use the instruments for monitoring as well as instructional improvement, and want to suggest four ways in which the monitoring instrument could be used for instructional improvement. Not directly, but indirectly.

One is, if we find there is a problem in the college graduates, that is worth addressing, then that will be advertised widely, and there will be a strong public movement to do something about it. Just as the *Nation at Risk* took the results of NAEP that were produced in the late 1970s and early 1980s, and used them to advertise the alleged deficiencies in the reasoning--among other things--of the students. And that had a tremendous impact.

A second thing that the monitoring/assessment device could provide is to be a model for the goals that many institutions might then adopt. When we set out the goals, and announce them loudly, this could serve as a model.

A third thing, is that the actual assessment procedures could be a model for assessment procedures that the higher education institutions could use. They could look at these assessment procedures and say "Hey, that's a good idea. Let's try something like that locally." And they could use it for local accountability, and they could use it for local experiments, research, and local feedback. But the thing is the administration would be different, so the states would be low in the monitoring.

Lastly, this monitoring instrument could be used in small research studies, which then would have feedback into our techniques for teaching at the higher level.

So, in summary, what I'm urging is a three-way distinction: monitoring, accountability, and improving teaching. And I'm suggesting that the test we use for monitoring not be used in the same administration for teaching improvement. Although it might be used in other administrations in other ways, for teaching improvement. They're both very important goals, but if we try to combine them in one administration of the test, then I think they'll both be defeated.

I'm Mark Weinstein from Montclair State, where I've been involved working with faculty from all disciplines, trying to infuse critical thinking in courses at all levels. And something came up that I think I want to use as an example for where consensus about generic skills could be seen to point, and--as John Daly's comment about problem solving indicates--it's clear that problem solving requires critical thinking. It's clear that critical thinkers ought to be adept at problem solving. It's also clear to people who know the tradition that problem solving has been developed within a discourse community that is far different from the discourse community within which critical thinking has been developed, and developed through engineers who have very different senses of how problem solving ought to be articulated, how it ought to be manifest, and how it ought to be measured.

For example, they use mechanical and technical problems, and don't use issues of political and cultural concern. Similarly, the generic skills represent truly universal areas of concern that all thoughtful people should be able to address in responsible ways. But how these areas of concern and how these skills are articulated, manifested, and assessed might look very, very different from the points of view of people who work in discourse frames as diverse as the physical sciences and the humanities.

And so what I would recommend is that not only should people like engineers engaged in problem solving have their say at what should be done, but people who accept the universal areas of concern identified with critical thinking--and maybe even the universal dispositions of mind that aid and abet critical thinking--people who have this legitimate concern but see that concern articulated through specific areas of study (natural sciences, social sciences, humanities, professional studies especially) be invited to report on what critical thinking these generic, universal skills look like when identified, articulated, manifested, and assessed within these special areas of concern.

I'm Dick Larson from the Lehman College of the City University of New York. And I come before this group, as I do before most groups, with a very strange combination of experiences. I am at the moment a professor of english. I have been a director of composition in a public university, I am working currently on developing approaches to incorporating writing into the academic disciplines. But long before that I was a faculty member for seven years at the Harvard School of Business Administration, teaching a course called "Written Analysis of Cases." And I enter this meeting with the conviction that, at the business school, teaching according to the case method, I was teaching a combination of CT, PS, and CM. In order to pass my course, the student had to do all of those things, and do them well.

The reason I take the floor is simply to say that it seems to me that there are profound possibilities in the assessment movement, related to Goal 5, for substantial improvement in undergraduate instruction, if the movement can lead to an awareness of how to approach faculty about changing their orientations toward undergraduate teaching. Making them more aware of the importance of undergraduate teaching. Making them more aware of the importance--as our group said yesterday--of helping the students learn, and indeed, "learn how to learn."

And my own conviction is that we have to infuse more widely into undergraduate curricula the acceptance of writing and thinking and CT and PS and also problem posing. As important elements in teaching processes and learning processes. In order to make that kind of infusion, we will, as Lorenz Boehm and no doubt others in this room are aware, we will have to engage in substantial efforts at faculty development, and helping faculty understand how they can do it; develop confidence that they can do it; and develop the recognition that by doing it they will in fact enhance their teaching--not detract from it.

So I see in the movement that this conference is initiating a very important chance to make improvements in undergraduate education, sort of along the lines that Ernest Boyer was talking about in his book (which I think came out last year) called *Scholarship Reconsidered*. Maybe we can, through this kind of movement, make scholarship include strength in teaching, and strength in assistance to students' learning.

I'm Don Lazere from Cal Poly in San Luis Obispo. I have heard very little emphasis throughout these meetings on the part of Goal 5 concerning every adult American having the knowledge and skills necessary to exercise the responsibilities of citizenship. I teach English Literature and Composition, so I'm not a political scientist. But I'm constantly overwhelmed in all of my writing and literature courses by the fact that whenever issues of civics, citizenship issues, come up the appalling level of student ignorance and indifference toward citizenship. So I would like to urge here that, in the future activities and projects of this project, that there be a strong emphasis on the application of CT, CM, and PS to the development of the rights and responsibilities of citizenship, and that when aspects and criteria for CT and so forth are defined, that there be a section defining and applying them to exactly what rights and responsibilities of citizenship need to be highlighted in reference to CT, CM, and PS. Maybe some political scientists might be brought into this effort, along the way, at that stage.

I'm Magda Kohlberg from the U.S. Office of Personnel Management. There are two things that perturb me somewhat about our conclusions today. One is the lack of attention to abstraction. We are talking about being sure that our assessment assesses CT and PS in real life situations. And this is well and good, but real life situations are always attached to contexts. And we have found--particularly in our job analysis of professional, administrative: higher level positions--that the capacity for abstraction and inferential potential in the ambiance of abstraction is extremely important. And I don't think that we can lose sight of this: it's very, very important.

The other thing is that somebody said--I think it was Group two but it may be Group One I don't want to take anyone's name in vain--it was important to define what a sound inference was. How do we know if an inference is sound, I think was the question. If I'm misquoting, please correct me. But I think (Group One was it?)

I think we must not lose sight of the fact that the soundness of inferences can be judged from their compatibility with logical schemas. I mean this is so, and should be something we keep in mind for the construction of this assessment. If we are to test, or to assess, inferential capacity (which is of course a skill within the CT domain), most definitely this compatibility with logical schematics should be kept in mind. I'm not saying that there is nothing beyond that. There is, in the creativity area, room for going beyond the bounds of the schema, but the compatibility with the schema must be there in order for an inference to be sound.

And even in the creativity area, in our research we are uncovering that creativity has a lot of schematics in it, too. Because it consists, to a large extent, in discovering connections that are implicit among phenomena, but have not been made explicit. So, the word of warning is: Let's not lose sight of the fact that sound inferences can be judged, as such.

That's all I have to say.

I'm Richard Paul. I want to sound a warning on the danger in conceiving the PS, CM, and CT in a narrow sense. Each of these areas can be viewed as a narrow specialty, or it can be viewed richly. It seems to me, in this context--in talking about the learning of students--that we want a rich understanding in each area, not a narrow, specialized understanding. For example, in my work, CT concerns originally emerged in a philosophical context. Philosophers were the main participants in the early conferences. And this had a natural tendency to be somewhat narrow, and to reflect the specialized interests of philosophers.

The field has since very much broadened out, and a rich concept is replacing what was earlier a narrow concept. The same is true of PS, and I think is true in the area of CM. One can look at these in terms of sort of the least common denominator, but one can also look at them richly, in terms of the way they interface across a wide variety of disciplines. It seems to me that--for the project we're concerned with--rich concepts in each area need to be what drives the test.

Now when you consider these richly--and here I would make an observation based on my own thinking, which you may or may not agree with--that they tend to converge. So when you try with a rich concept of CT to distinguish it from effective PS, you have great difficulty. Because if you've got somebody you call a very good critical thinker who's not very effective in solving problems, you have a virtual contradiction in terms.

If you consider a critical thinker simply as a "critiquer" of the products of others like an evaluator at the end of a certain process, then a literary critic may not be a good poet or a novelist, and so forth. And then you view CT as simply being a critic. But that's a narrow sense of CT, it's not the sense in which we want CT across the curriculum. So, it seems to me you want to be very careful to bring in those people in the areas, who approach the areas, richly and broadly, with a sense of interdisciplinariness, and not those who speak for the area in a very narrow, specialized way.

And I think this is a very important thing which will bear on the credibility, and the usefulness, of the assessment instrument that emerges. Also a comment on the question of abstractness, putting things in context. The kinds of problems that we should assess, that are real world problems, are not ones that are so fixed in a particular context that they are idiosyncratic, but rather those problems that are real world problems that are broad and cross discipline areas. Lots of problems of ecology, for example, involve reflecting in an historical and a political and an economic and a moral sense on the same problem. That is, the problem has many dimensions to it. It is also embedded in a variety of contexts. And this kind of thinking then involves PS, involves the use of language in very effective ways, involves CT, and undoubtedly involves background information and other kinds of considerations, which may or may not be put into the prompt itself.

Finally, the point about inference was really a point about intellectual standards in general. One way to understand CT is a concern with the intellectual standard, so that as students learn to reason historically, learn to reason economically, learn to reason mathematically, learn to reason scientifically, they should come out with intellectual standards which they use for the purpose of assessing their own thinking, both in that domain, and beyond that domain. And this is integral to our understanding of what CT is. In this case it is to be distinguished from simply descriptions of thinkers, from descriptions of expert thinkers, descriptions of novice thinkers. We're talking about the kinds of standards that students should come out with. Intellectual standards at the end of their college career. It is my observation that if you ask most graduating seniors, "What intellectual standards have you learned, that you now hold your thinking responsible for? You would find that students would draw a blank. That is, present instruction does not call attention to intellectual standards. It tends to be heavily focused on content, and the re-iteration of content in lectures and textbooks. So I think there's a substantial problem here. And if you understand CT is connected with intellectual standards, you see it in a somewhat different way, and I think a richer way.

Michael Scriven from Pacific and Western Michigan. Three quick points. It's kind of implicit in most of what's been said--but hasn't been made explicit--that we want to be very careful to stay with the english vocabulary, in talking about CT. When I first started teaching logic, in 1952, at the University of Minnesota, I heard, to my surprise, that somebody was teaching logic on the Ag(riculture) campus in an extremely scientific way. And had managed to demonstrate 400% pre-post gains. And I thought that this was something I had better learn from, so I trekked over to the AG campus, and it was in fact true. So I got to see the pre-test, and it said "in the syllogism in the mood sorites, is there a distributed middle premise or isn't there?" Well, on that sort of stuff, you can get 400% gains pretty easily.

Well we want to watch ourselves a bit with this, and make clear that there are roughly 72 words in the english language which are in our common vocabulary, which are terms of logical or critical appraisal, and that's a pretty good, rich vocabulary, and we want to stay with it as much as we can.

Second, I think that there's a format that's emerging from the discussion which I want to utter a caution about. The format is: there should be the general CT section of the master test, whatever it is, and then there will be the subject matter-specific sections. And this is our kind of bowing in the direction of the importance of CT in the disciplines, which is indeed an important matter. But I want to try to push rather hard for not setting the test up so that you get an option of your choice of interpreting poetry, or doing analysis of thermodynamic phenomena in the second part, but I want to make sure that massive extra credit is available if you can do them all.

That is, I think that the contribution of good CT instruction to PS in particular but to CT in general, is mastering the general methodology of half a dozen general areas. I think we all ought to be literate with respect to the notion of social science control groups, to the notion of lab standards, and measurement procedures, and observable errors, and so on and so on. It's not that hard, but it's something we won't all master at the age of sixteen. But it's something we ought to keep working towards. I think it's important not to assume that there isn't a reasonable part of PS and CT which involves mastering a large number of methodologies. And so, there should be the option to do as many as you can, or part III, and get extra credit for having done so, if you get the answers right, of course.

One comment about what Bob Ennis was saying. Something important to us in our group, which I didn't hear Jean emphasizing in the summary, was that we started (at least I started) with the feeling that matrix sampling was going to be the way to go to handle costs, and so on. But I got persuaded by the other people in the group that that won't really do. But it doesn't lead you into the problems that Bob was warning us about. In matrix sampling, as he rightly pointed out, you can't give a very enlightening feedback to the individual.

But if you go for full tests for each of your sample, but do not undertake to take a large enough sample from institutions--or for that matter from states--so that you can give a report at the institutional or state level, you don't get into high stakes. What you do get is really important: you get an incentive to participate. Which is something we've got to take extremely seriously from Day One. And the incentive is "here are some really important skills. If you can give us a certain amount of your time, we'll give you feedback on your performance on those skills, **and** we'll give you a certified transcript which you may--at your option--use in applying for jobs."

Now there's no need at all for that to be treated as high stakes for institutions, for instructors, etc. But we should, I think, make the tests available in some format--a parallel form of them--so that instructors, and for that matter institutions that wish to participate in having some institutional measurement made, can do so, too.

My name is Ted Marchesi. These comments are so good, that I find myself forgetting what I was going to say twenty minutes ago. The last seven comments have been superb. My intent was actually to make some remarks directed at the OERI staff, as much as at my colleagues. And that was to compare and contrast, if you will, a little between the reports from Group Two (which I was a member of) and that of the other three groups. If I had a friendly chide of the other groups--which consisted, by my impression, primarily of educational researchers and specialists in the abilities--it is that they found no good idea that they didn't embrace. And Group Three especially was in favor of everything, every purpose, every end, every program, every student every method, everything altogether. And Group One had a research agenda that wouldn't quit. If you had ten years and a hundred million dollars, you could do all of that to the hearty cheers, I am sure, of everybody in the room.

Now Group Two consisted primarily of practitioners. We're very aware of how pressed many of the on-campus efforts that Peg Miller--who was in our group--are. Over the last six years we've learned a lot about the doing of assessment by faculty that is directly related to the improvement of teaching and learning. What we've also found out in the last six years is that we don't know how to answer the public's question, which is embodied in Goal 5.5, and that is, "What is your contribution to student learning?" "With respect to these three abilities, what do your graduates know, and can they do what your degrees imply.?" We can't answer the public's questions.

We also feel that the time that we're going to have to answer these questions is not ten but perhaps two years, maybe three. That we're never going to have a hundred million dollars, we might have one or three million. And that we need to do the best focused thing in the time immediately ahead to teach ourselves how to take the experience and the knowledge that we already have, and devise ways of answering the public's questions about our contributions to student learning. And that is what the proposal of Group Two consists of.

Now all of the other things are nice. International comparisons, feedbacks to every single student and program and professor, and curriculum, and major, and institution, and everything else like that. That's all wonderful and desirable in many ways, but the essence is to answer the public's question, and not to push aside--or even nationalize--assessment and all the things that are going on campus.

Now does this mean we have no ideas whatsoever about the improvement of teaching and learning? No indeed. I'm pointing out, first of all, that there are already a rich array of things going on a huge number of campuses, that are related to assessment, that are related to teaching and learning. But the special contribution to that conversation that this goal-directed effort could make is what, actually, Professor Ennis described in the first of these comments. That is, this effort ought to put CT, PS, and CM abilities into the public and the institutional and faculty minds in a way that it is not now.

Now that's the very important kind of thing. If it makes faculties, and institutions collectively aware that these are important things that should happen in undergraduate education, and raise demand within them for ways of arranging curricula and pedagogy so that they're more likely to occur, that's the larger outcome, rather than particular feedback to me as a teacher of sophomore organic chemistry, on how to fix my course. It's not that. The feedback that we want is that we want to put these three things more firmly in the public mind, and in the faculty mind, and have programs come behind that tell people how they can more likely achieve these kinds of outcomes.

But we want something that is much more focused and doable. That takes what we know and helps us answer a very important, pressing question.

I'm Ellie Greenberg from Denver Colorado. I think by way of introduction, I have this odd combination of professional experience, that walks between having designed programs, and being a researcher, and most recently, being a designer for a corporation, in workforce development program for fourteen states, that involved about 1,400 institutions of varying sorts. And so I bring my thinking to this meeting out of those experiences. And I guess I just wanted to put somewhere on the agenda, which I don't think occurred in our group, a couple of things, without discussing them in great detail.

One is what we've learned about gender differences in learning. We have a body of research that is important, relative to how men and women learn, and there are similarities and differences. Many of you are familiar with that research. I would like it not to be lost in this discussion.

Number Two, there is another body of research that is rich, and is concerned with thinking about thinking. And many of you are familiar with that. And it is referred to as developmental research. And it is less about disciplines than it is about the quality, and the complexity, of thought. And where our understandings come from, and how we make judgements, which obviously relates to citizenship, and those kinds of language that are more common in the public arena than the academic arena. We don't talk about citizenship a whole lot in the Academy. But the public does talk about it, because it sort of grabs the whole idea of what we are as a nation.

So I think the developmental research informs how it is we construct this notion of assessment. That's learner-driven language, and the third element of that is my discomfort about making "lifelong learning" or "learning how to learn" a separate matter. And how age, somehow, appears not to be an issue, and the cyclical nature of people using schools over their lifetime is relevant to this discussion.

We do not simply enter, middle, and go in four years, in the old pattern. We all know that, and we need to pay attention to our own data bout that, and who these learners are, and how they go through what I call the good revolving door. They come and they go, and they come and they go. And they're stopping in and stopping out, and stopping in, and stopping out. And this is a pattern that we need to affirm, that we need to understand, and we need to intervene, perhaps, on occasion, and say "Hey, how has it worked for you?" And we might pull those moments, the moments of assessment, and follow those persons over time, and I think in many ways that's quite affordable and doable, and in fact if we don't do it we will have a very narrow snapshot. And I don't think that's what the intent of the goal structure is.

The decade-long 2000 focus, the 2020 focused adventure, which we are now given the opportunity to join. And I think that's really very powerful. So, longitudinal looks at how people learn through life and use institutions to do that as they go, it seems to me, is the pattern that's being spoken of.

That leads to a fourth thing I suspect, that in our group unfortunately we didn't have enough time to get into. But it happens to be a lot of fun to get into, as well as a lot of aggravation. And that is the whole structure of the thing, and how we proceed to make it happen. And I guess that I feel sort of deprived of that conversation in the political sense. Where do states fit? Moving from the learner out of the institutions, which has been the focus of discussion, into the societal question. Who are the stakeholders? Why was the question asked? and Who is the client, or clients, as the case may be.

And therefor, how the federal department can harness and play consultant to the nation in this matter. And in some way create and support the kind of network organization that is embedded. Such as different ways of assessing. Which are clearly articulated in our materials by our Alverno friends, and other ways of assessing, and how to make those parts of the pie important in this effort, so that the diversity of how we do it is somehow captured by our federal agency in support of that diversity. And I think this is a real opportunity for the role of the department, and NCES in particular, to shift, and to become a model for collaboration. In the way in which institutions don't typically do, but that when you have the lever, and you can't control the nodes in the system--which you can't in this instance--then one must figure out how to manage the network organization in a very different way. That's all I have time to say at this point, but I'm sorry we didn't have a chance to discuss the structural and political implications of the task.

Norman Fredericksen, ETS. How many cognitive psychologists in the audience? None, or one hand went up a little bit. How many know what a schemata is? (Show of hands) Good. I won't have to speak so much. Well for those who don't understand, I'll just say that a schema in the language of the cognitive psychologist is a cluster involving pieces of knowledge and skills and so on that are closely related. One part of it can't exist without the other part, as time goes on.

An example in the literature refers to restaurant schemata, which consists of deciding on a restaurant, going in, waiting to be seated, being handed a menu, you make up your mind, and so on, and so on, you know the rest of it: a cluster of closely related things that people do, and expect to do, and are not related to a lot of other dissimilar schemas. Schemata is the plural, but I'm not sure many people use it anymore.

So when we think about what goes on in training, in kindergarten, you don't have separate classes with the different aspects of learning. But as the higher grades come along, people are separated, and they teach arithmetic in one room, and language in another room, and things get more complicated, and you get into college, actually when you get into graduate school, these schemas shrink, or disappear. Think of college course, any one course such as a math course, begins with very simple things, and as time goes on and assignments are given, you develop a much larger and richer set of related ideas and skills.

I can't think of CT, CM and PS as one huge schema. That's all I want to say.

Sal Corallo: Open mike?

My name is Mike Knight. I thought he said open, Mike, that's why I came up. I have a number of concerns, but I am heartened and worried at the same time. When I hear folks use the phrase "the test," that worries me. But I've resolved that problem. Because I'm going to design the test. I'll send it to you. I know that all of you will accept it, and use it. Are there any disagreements on that?

I think I've just described a much more complex process that will really describe what goes on campuses. If I were to do that, I can imagine how long you would discuss the test. Then you would discuss me. Then you would discuss my parents, and you would raise certain questions about my lineage.

Now I say that because this describes what our Philosophy Department has done. Our Philosophy Department has said that assessment is impossible. If it were possible, it would be worthless. In the same meeting, they went on to discuss the President of the college, the governor, and the chancellor, and this discussion focused, again, on lineage.

They have moved forward. And they have designed what I believe and think is a very critical, very rigorous assessment process for philosophers. I will point out also that I do not believe philosophers are that different from any other discipline. Any of my colleagues that would not have questions, critically developed questions about assessment would not please me. It's not their job to please me. But I want my colleagues to be critical, just as I've heard all of you being critical today.

A concern. When I hear someone describe assessment that will lead to improvement, that concerns me. Because assessment includes improvement. If you do not think that improvement is a part of assessment. I would ask you to re-think it. Peg described the circumstance about using results. Our quintessential question with assessment is who will do what with these data?

Because we don't want to collect stuff that we're not going to use. Action, action, action.

Now this was not easily arrived at. I can describe it, what happened, very simply. In our initial steps over the first eighteen months, people began to express concerns that data were piling up. In my mind it was so clear that action was the final step to close the loop, that myself and my colleague Don Rumsen did not communicate it effectively. Improvement is a part of assessment. I also have a view that I believe is different. I believe, again, Peg mentioned it, I believe Ed White mentioned it.

I would like to raise the question to you, How will you convince the people on your campus that there is no bad news. If you can convince people of that fact, then you have resolved most of the major issue of organizational change. How will you convince them? How will you create an environment where that is acceptable? Now if the belief is that if I reveal bad news I will suffer, the consequences are obvious. I think you can figure that out for yourself.

I will just tell you one story that I think is amusing and informative. It's called the paranoia shift story. Someone told me I was paranoid, I said, "I was afraid of that." No, no, this is a somewhat more serious story. When we began assessment on our campus, there was resistance. We anticipated resistance. We would have been astonished if there hadn't been resistance. This is the way we conceptualized this project. When we first discussed it, it was described as a student development project. Then, a curriculum development project. Then a faculty development project. And it is all of those, but it is more than that. Peg, again, mentioned reward structure. It is an organizational change project. If it is not seen as an organizational change project, I do not believe it will be successful.

This the paranoia shift story. One of our psychologists, about six months into the actual work of assessment, reported at one of our faculty Senate meetings, that he had observed a paranoia shift. The paranoia shift was "Why must I do this?" to "Why have I been excluded from doing this?" Now I do not expect my enthusiasm to be infectious. A number of the people on our campus are very excited about assessment. I think I'll end with that.

I'm Peter Ewell, a member of the irrepressible Group Two. I think that by the time this is over, you will have heard from all of us individually. (Laughter and applause.) The remarkable thing about it is that we did agree completely with everything that Ron Hall had to say about it. We are a practical group. We are a group that has had a fair amount of experience in watching states wrestle with this issue for about the last five or six years, and watching institutions wrestle with this issue.

I think that I can speak--the group will probably disagree with me--but I'll attempt to speak for the group in saying we think this is a noble enterprise. We are happy to be a part of it. We are honored to be a part of it. We think it's a conversation that has to happen. At the same time I think that we would be remiss as critical thinkers--which seems to be the center of gravity about all of this--if we didn't raise a couple of questions about some basic assumptions of the enterprise, again reflected in some of the comments of my colleagues.

But let me mention two of them. The first assumption behind all of this is that the Goal is about improvement, and that our part--in this room--of the enterprise is to try to arrive at some ways of detecting improvement, and informing improvement. Absent the mechanism for improvement, this is a useless exercise. I think that several people have said that, but I think that it has to be constantly raised. And throughout the instrument- or device-development process, we have to constantly ask the question. Are the last of the pieces in place? Is there evidence that the rest of the pieces are in place? How is what we are doing connected to anything, in terms of a mechanism for change?

The second assumption is that we're going--and I think Ted's comments pointed this out--is that we're going to have infinite time and money to do this. I think that the major difficulty that you can see in the experience of states in trying to (although I'm out of my field on this one, I think I can generalize to the case a little bit here) but I can certainly speak for the efforts of assessment in higher education. That you have to look at whatever program you design, as though it were half implemented. You have to look at it as though the wonderful thing you've put together is going to have to be implemented with half the resources that you expected, in half the time you expected, with a great deal of political interference, with a great deal of special interest lobbying, with a great deal of modifications to that design, that were not taken into account at the beginning.

As a result of that, I want to revive and stretch a little bit the concept (people have been talking about stretching the concept of validity) of robustness. I think that what we've got to be doing is designing a set of instruments, approaches, and devices which are very robust. They are things which will give us some information in spite of the difficulties of half implementation and political interference. That's not an easy task. But I think it's definitely one that we should be paying a good deal of attention to.

Richard Paul--A very brief comment with respect to the quality of instruction at postsecondary institutions. Allen Schoenfeld, who is the distinguished mathematician at Berkeley and engaged in research into math instruction at all levels, recounts the following story about elementary school math instruction.

Children are asked the question: "There are 75 sheep in the field, and five sheep dogs. How old is the shepherd?" Four out of five students add, subtract, multiply and divide in order to compute the age of the shepherd from the number of sheep and sheep dogs in the field. And the more math they've had, the greater this tendency. Now Allen Schoenfeld studies calculus instruction and advanced mathematics instruction at the post-secondary level, and he takes students from a university (I forget now which) advanced calculus program, and gives them a simple algebra question at the end of the course, and finds that only 20 percent can do it.

He takes his senior math majors at Berkeley and gives them a tenth grade geometry problem on an advanced examination, and finds that a low number of students can do it, and most try to use advanced math to solve a problem which can be solved by very simple geometry. Now he generalizes about this, in a book on mathematics problem solving in the following way. He says most math instruction, both at the pre- and post-secondary levels, has two characteristics. One, they're deceptive; and two, they're fraudulent.

His basic conclusion is that students do not learn to think mathematically in math classes, and he and the National Council of Teachers and Mathematicians, and also the American Mathematical Society and the National Academy of Sciences are concerned for the quality of math instruction at all levels, because they recognize there's a pattern of all levels that consists of the following.

One, an algorithm is introduced. Two, the instructor illustrates the algorithm in front of the class. Three, the students practice using the algorithm. And four, the students are tested on that algorithm with standard questions and problems that are quite like the problems the professor used. Allen Schoenfeld says this is a perfect design to produce non-mathematical thinking. Now, there is empirical research in all of the disciplines that parallels this research that is going on in mathematics instruction.

We have a fundamental problem at the post-secondary level that can be empirically demonstrated. I'm out at two or three campuses every month, working with faculty on these kinds of problems. There are many faculty members who are absolutely addicted to didactic instruction. What they believe in is coverage. What they believe in is lecturing. And what they believe in--though they don't know it--is rote memorization. And they believe that throwing a lot of stuff at the students, through lecture, is really the way to get people to end up as good reasoners in mathematics and history, and so forth and so on.

So there is a fundamental problem. It can be empirically demonstrated. And I hope this process contributes to the remediation of that problem.

I'm Joan Mills, I'm a member of Group Three. I just felt a need to come up and speak in behalf of some of the other groups, that there was some practical, seasoned people in other groups. I say that a little facetiously, but I do respect, indeed, some of the concerns that came out of Group Two, but to suggest that the recommendations that came from the other groups could in fact not be accommodated I think would be inappropriate.

I think there have been a couple of observations that are fairly important for us to remember. And part of that goes to the fact that, in terms of the context of the National Goals exercise, we really are talking about assessment of consequences. And it also needs, then, to inform and improvement agenda. And clearly, in Group Two are the people who have to live with these kinds of issues every day, were recognizing, and groping with that tension constantly.

As I listened this morning, it seems to me that we have some really clear and wonderful opportunities, to think very differently--and indeed learn from the lessons of NAEP and other kinds of exercises that we have gone through in the past. At least in terms of Group Three, when we were talking for example, about a much larger survey, dealing with a larger population, we were recognizing, indeed, that people come and go, in and out of the postsecondary institutions, and that we in fact need some baseline data so that we can begin to understand what an assessment of consequences means.

So we were not assuming that Emerson was going to be able to raise an incredible amount of money, that all of us who know anything about Washington D.C. know won't be possible, and we also think that this agenda needs to fit in to other activities that are going on in other parts of the federal government, so that we can begin to mount and create an assessment of consequences in the most informed way possible. A lot of that has to do with work that is taking place in the Department of Labor.

As I listened this morning, it occurred to me that there are several lessons from NAEP that as we begin to develop this design, that it won't have to be quite as standoffish, as we had for NAEP, over along number of years with the educational institutions. And indeed there are some wonderful lessons, as NAEP has been expanded at the K through 12 level. And I really want to throw out a challenge, in fact to some of the national institutions that are sitting here. There are some interesting kinds of models.

For example, the Council of Chief State School Officers have developed an assessment center that is working hand in glove with--not always agreeing with what goes on in terms of what the federal government is doing with NAEP--but it is a way to begin to develop an agenda that has an improvement agenda as well as a consequences and assessment agenda. I think a lot of the organizations who are out there in the audience today can in fact make those kinds of things begin to happen.

I really would urge us to remember that there are a lot of other organizations that indeed do, in fact, need to be involved. For example, in Group Three, when we talked about proprietary institutions, two-year colleges and etc., those weren't accidents. This is not just a four-year college concern that we have. And so we need to make very clear that we reach all of those.

And in fact created a line that involves lots of different kinds of institutions, but not confuse the issue of an assessment for consequences as being a part of the major reason that this ever got into the Goals to begin with.

CLOSING EXERCISE
COMPILATION OF RESPONSES TO QUESTIONS

Listed below are the direct responses from 27 participants who attended the study design workshop. Respondents are not identified. However each questionnaire has been coded by alphabet, so that it is possible to group responses to each question by questionnaire. That is (A) in questions 1-3 all came from the same questionnaire.

QUESTION 1 :WHAT WOULD YOU LIKE TO SEE RESULT FROM THIS EFFORT TO DEVELOP STRATEGIES FOR ASSESSING HIGHER ORDER THINKING AND COMMUNICATION SKILLS OF COLLEGE GRADUATES?

We need a better understanding of the state of peoples ability in critical thinking communication and problem solving and how they relate to work success, success in graduate education, and the quality of life. (A)

Are higher order thinking and communication skills important for strong emphasis in higher education or are they only a mirage, the true indicators being knowledge in content area (or some combination of both)? (A)

A national system for monitoring progress toward goal and an effective system for communicating results, encouraging reflection and action (not institution-bashing). (B)

A parallel strategy, including incentives to encourage individual institutions to analyze their effectiveness and to strengthen their programs accordingly. (Assessment is only one part of the strategy.) (B)

The process of developing those strategies (of assessment) can have an important disciplinary consequences in terms of defining concepts, developing instrumentation and even determining what counts.... For the first time we would have; 1, a barometer of America's communication contingencies; 2, an indicator of critical communication competencies; and 3, a good sense of how communication affects other variables. (C)

An increased focus on critical thinking, etc., as the outcome of education. An assessment that drives reflection on the goals and methods of instruction. (D)

Group Two's notion of institution based assessments gives a qualitatively different approach to the notion of a national assessment. (I think it is similar to the model programs of Project Head Start.) I think this approach is innovative and allows us to combine issues of "accountability" and "informing instruction." (E)

A fundamental change in the instructional process--faculty development--Most faculty don't teach critical thinking skills--until instruction changes--I don't see much progress. (F)

More collaboration with K-12 and the business community on solving the problem. (F)

A focussed research agenda on national assessment for postsecondary education. Establishment of some advisory panels technical, substantive, etc.) to inform the design of initial efforts. (G)

I would like to see a national indicator of such skills, baseline data gathered before the entire program gets underway. And then the state and institutional data available for those who want it. (H)

I would like to see assessment and results left to the national level (like NAEP) and the improvement interventions left to states and institutions. (H)

I would like to see something realistic come out of all of thisnot some "pie in the sky", lets leave it up to the institutions. That (leaving it to the institutions) will never work! (H)

Exemplars of excellent, successful models of education in colleges and universities. Many kinds. (I)

Public discussion of evidence for a good college education. (I)

Improvement in educational programs of higher education institutions. (I)

Better understanding by researchers of relationship between critical thinking and domain knowledge. (I)

Clearer definition of critical thinking and communication skills (J)

Greater role for Federal Government in Higher Education (J)

The awakening of higher education presidents that measures of quality are needed. (J)

Something like Group Two's institution based development process. (K)

As a college teacher of English, I would like to see more government based support for instructional related research at the college level on factors promoting or impeding higher order thinking and communication skills. (L)

I would not like to see a process leading to a test or a national indicator. I would like to see more examination of existing procedures, proxy measures, and research and evaluation of possible instruments. (M)

Blank (N)

Provide some funds for institutions to focus upon improvement efforts (faculty development, using intensive focused teaching strategies dealing with individual student learning styles, etc.) that will use multiple assessment strategies in establishing their worth. (ED note, this seems to suggest faculty recognize "individual learning styles" and adjust their teaching to address these individual variations.) This is a focus on the literal interpretation of Objective 5.5--on improving CT, CS, PS, rather than on assessment per se. Using the assessment will show (or not), via pre- and post-treatment use. (O)

I am surprised that indirect, non-obtrusive measures/indicators were not discussed more. Why not survey employers about the quality and skills of graduates, or survey deans of graduate schools for other indicators of progress? Also why wasn't employer based education addressed? (P)

Fundamental principles for the assessment of CT, PS, CS. A national measure steeped in the "voodoo" of psychometrics but having little applicability will gain little acceptance. Moreover, somehow faculty/departments/institutions must be able to pursue assessment of CT,PS, CS on their own for the improvement of teaching and learning. (Q)

1. Reform of instruction; 2. public articulation of a rich concept of the significance of critical thinking and communication skills, that enable the public to grasp their significance and the extent to which our schooling has failed to cultivate them; and 3. an excellent national assessment process in these vital areas. (R)

Blank (S)

A valid quality assessment program with the ultimate goal of improving instruction. It is vital that the assessment be performance based and strongly grounded in principles developed by critical thinking and communications theoreticians and researchers. (T).

A strategy (including a set of indicators and instruments) to help move along and monitor progress in meeting both the political agenda (information and improvement) and the underlying educational agendas. (U)

A carefully thought through process for development involving a wider conversation among parties-of-interest in this process--institutions, state leaders, and methodologists. It should be repeated on a wider scale as the assessment development process proceeds. (V)

The development of a National Network Organizational System. That can be used throughout postsecondary education (all kinds of providers) to learn about how and what learners of all ages are learning, over time. Whatever the particular "skill" or "competency" that is the focus of our inquiries. This is a process. This is an opportunity to change the way we do things in recognition of learning throughout life. (W)

One that provides , every five to ten years, an assessment of individuals who are entering and involved in the workplace--using the same set of instruments that are used for students. Hard and expensive but DOL should be encouraged to be a joint funder. There is no reason this type of assessment can't provide double social utility but there is also no need to conduct the assessment as frequently as would be done on campuses. (X)

A. Not an overly deliberate, but a concerted research plan focused primarily on validity issues. Many have suggested instruments or approaches that have not been thoroughly investigated, or consequences of those instruments that haven't been fully examined. A national assessment should be well founded in research and validated. (Y)

B. Accurate data! May be hard to get particularly with the politics of higher education to be essential if the assessment effort is not to be wasted. (Y)

A valid usable system for responding to the national goal but primarily focussed on the teaching-learning process. If this is not accomplished we will just have another data gathering mechanism the won't facilitate improvement which is a key word in the goal statement. (Z)

BLANK (AA)

QUESTION 2. WHAT SHOULD BE THE NEXT STEP IN THIS PROCESS?

Clearly some plan of action is needed. This could be a map of a strategy over the next five years. Part of the this plan of action would be the creation of a model that would depict how higher education experiences develop CT, PS, and CS and how these relate to effectiveness in business life after college. A series of short studies need to be conducted to provide information on major questions; when to assess "college graduates", how to know what is enough. (A)

1. Consider design indirect measures as interim strategy to give sufficient time for R&D on direct measures and associated issues. (motivation and standards). Validity of new measures is essential and will require sustained research. Equity and impact on minorities also requires attention. (B)

2. Convene consensus process to specify and prioritize skill domains. (B)

1. Through reviews of what we know about the three areas in terms of assessment and validity. I was impressed that there is so much confusion in the critical thinking area about the dimensions and empirical properties underlying them. Maybe separate conferences in each of the three topics; CT, PS, & CS. (C)

2. Preliminary(or even hypothetical) development of instrumentation to get a feel of some of the issues and problems associated with the project. I find it useful to "try out' an instrument. It gets a focused reaction. (C)

3. More papers on communications and problem solving. So far the emphasis has been on critical thinking. people are assuming this is a critical thinking project. I don't think this is appropriate. (C)

A working group organized , either on site here or in selected sites elsewhere, to articulate a rich compendium of CT, PS, and CS relevant to college learning and its useful application, (D)

(Identify)... the link among the conception and the abilities, the measurement of them and improving instruction/student learning needs work. (E)

A. Find out if any college/university etc., (other than Alverno, Kean, UT (Tennessee)0 has used critical thinking assessment to improve instruction. (F)

B. You need to "plant" several articles in national publication--begin to orient faculty to the reality of assessment.

C. Keep dialogue going with this conference's participants. Using "Delphi Technique", keep us involved. We have devoted time and thought to this--Keep us motivated through good communication. Maybe in the spring and summer--we need a retreat--away from everything--to put together a proposal--I agree with Ted Marchese--focus upon what is doable--Answer question--what do we contribute to student learning? I also agree with the comment "What is Assessment going to tell us that we don't already know about students? These are questions that we need to answer--good aspects to cover in a "Delphi Approach" as a follow up. We just scratched the surface of this issue--How can we harness the energy and questions we have all unleashed? (E)

Meet with a smaller group of external people to discuss research/design/development issues. (G)

* Clearly set forth the purpose (long and short term) of such an effort. (H)

* Give some idea of the kind of funding that might be available to support the effort. (H)

* Operationally define "problem solving", and "critical thinking". Identify specific skills for each and get to work on an instrument(s) to measure them so that a baseline can be obtained. (H)

* Get states and institutions to research/tryout various interventions. (H)

* Research to find out "when" these skills are learned or developed (It might be long before college). (H)

Development of a five year plan by NCES, with long and short term goals (I)

Get this group back together to react to your synthesis of this workshop (J)

Continue to include practitioners in the design process. (J)

Please very early on, initiate a search for an analytically useful conceptual model or models of critical thinking. Simply narrowing the field will produce benefits in instrument design, utility, and integration of measurement across such diverse domains as communications, critical thinking, and problem solving. (K)

OERI funding of such research (L)

Group 2 recommendations. I agree with Emerson that it should be more "research than survey like" (M)

Tests even crude first versions; several versions. (N)

Small grants to get alternative versions of the of the test field tested (N)

NCES Synthesizes results into a single version (N)

Run on a large scale (N)

Secure funds for the R&D effort described in #1 above and get out an RFP. Involve FIPSE Staff in developing the RFP because they may know many of the good practices in higher education that need more trial/further dissemination. (O)

NCES should consult with other Department of Education offices (and other Federal Departments, eg, Labor) to see what is already known and what has already been funded. For example, what do we know from FIPSE and National R&D Centers in OR?OERI (Education and Quality of the Workforce. Then go outside the education community. (P)

Systemic investigation of measurement type feasibility, creditability, and exportability coupled with a continuation of the present multiple constituency/expertise conversation. (Q)

One thing that should be done is the funding of some pilot items (prompts,...) that embody rich concepts of critical thinking, problem solving, and communication skills, especially those which successfully provide evidence in multiple directions. (eg. performances that are simultaneously illustrative of critical thinking abilities, writing/listening abilities, and problem solving abilities. (R)

Development of a plan for R&D and preliminary design. The plan should include long-range (and) strategic vision. (S)
Set up core content sub-groups and methodology groups to investigate domain & develop suggestions or guidelines for proceeding. ("What needs to be done?" How do we go about doing it?) These subgroups must be from the content disciplines to have credibility and validity. Perhaps they could meet with psychometricians. (T)

Preparation of several draft approaches for comment and criticism--More communication with users in the field--Articulate clearly the purposes and objectives, but differentiate NCES roles from those responsible for actual improvements in practice. (U)

Proceed with designing a development process for a demonstration based upon the recommendations of Group 2--an Institution centered approach, guided by several alternative assessment designs.

1. Connect with Department of Labor's efforts, CSSO, SHEEO's, Lav's, etc. and make a coherent, understandable statement about the partnership system being created to improve lifelong education in America. (W)

2. Convene small workgroup too create the structural design (savvy and nitty gritty).

(Note: see also comments of W above in question 1) (W)

3. Meet jointly -Lav's--to hear how the who did/created the tools--conceived the goal statement and what they want as a result. Consider who the clients are. (W)

I thought Gary and Emerson's comments on next step were generally on target, ie. reaching out to broaden networks, beginning with some R&D work--with one exception. Admittedly the concern is not well thought out but I had a negative reaction to the possibility to establishing a "Commission", ala the clearances et al NAEP one of a few years ago. The reason being that at least for the moment we seem to be drawing on Commission's and Task Forces and it is not hard the imagine burying issues. Goals Panel et al. At that point NCES wasn't trusted--today it is. Maybe asking NAS to convene a panel would be safer--just a thought. (X)

Funded research (Y)

Continued involvement of all significant constituencies (Y)

Combine approaches of Group 1 & 2. Group @ has a greater sense of reality. Hopefully existing instruments can be used. I would hate to see this exercise become a boondoggle for test developers. States need to be drawn into the conversation. Be cost effective. Existing instruments may not be perfect but the would added cost of developing new ones be worth the millions? (Z)

1 Using existing instruments that assess the proper constructs--begin to see what's the state of the current situation, EG, Use the California Critical Thinking Skills Test to take a broad--but not deep survey of Delphi (APA) Report's core CT skills. (AA)

Start R&D on suitable assessment strategies to target the proper domains in depth and breadth. (AA)

WHAT DO YOU SEE AS THE MAJOR PROBLEM(S) OR CONCERN(S) IN DEVELOPING THE PROCESS AND/OR CONTENT OF THIS ACTIVITY?

One major problem is that the project will serve to conform college experiences. There needs to be some protection built into the project to repeat and maintain diversity. I do not believe that critical thinking, communication skills, and problem solving is manifest in the same way in all fields. An artist who graduated from colleges from a college may be a very critical thinker in communicating through an art form that would not be identified on all instruments. Somehow this ability should be valued. How to value diversity (A)

1. Motivation - Enticing students to participate and give their best effort, if direct measures are used. (B)

2. Reaching consensus on appropriate skills for assessment--given the diversity of institutions whose students are of interest--without concentrating on "minimum" skills and leaving out the best of capability. (B)

3. Unintended negative outcomes of the assessment process. Plan now for a study of the consequences and impact of the assessment program. (B)

1. Validates - Do the measures tap into things that have real consequences outside the college. (C)

2. Expense - performance based measures are quite expensive. (C)

3. Cultural Diversity - Each of these areas are quite open to critiques about assumptions. (C)

4. Outcomes of the Process - It is quite possible that someone could reason poorly and still come out with a highly sophisticated solution--if solution are topped we may grade without process. (C)

5. Defining "advanced" and "effectively". (C)

Drawing together working teams that reflect the range of relevant concerns. (D)

Move away from "the four questions" to a more in depth discussion of how to integrate accountability and informing instruction--focus upon how to improve instruction and find ways that institutions and faculty can feel they have a stake in it--that it's not top down or external but integrated into the institution. (E)

1. Trying to "do too much." (F)

2. Focussing on "exam" over "learning process" ("symptom" over "Illness"). (F)

3. Resistance at faculty and administrative levels. (F)

4. Not enough dollars to do it well--wise use of dollars available. (F)

5. May focus on "Quality" over "Access" - What will this mean to the "open door"? (F)

Acceptance by institutions. Understanding by the public. (G)

* Overcome suspicions by states and institutions (especially faculty.) (H)

* Motivating students to not only "take", but do their best, on whatever assessment is decided upon. (H)

* Managing a program nationally, but deeply involving the states/institutions. (H)

* Arriving at consensus about what higher order thinking is and how to measure it. (H)

- Mobilizing the relevant communities in higher education, including the disciplinary associations. (I)

- Developing a content...(instrument?)... for what is being measured, in a form that the public can understand. (I)

- Appropriate instrumentation. (I)

- Adequate research design for a development project. (I)

1. Money to support performance such as communication (Interpersonal skills). (J)

2. Motivation of students and institutions to participate seriously, start working on college presidents now. (J)

3. Problems of standard setting. Current methods for cutoff score setting are too arbitrary, (J)

4. Whether to assess non-college persons. (J)

Courage to resist the press toward a quick, single nationally responsive instrument. We simply do not, as a profession, have an adequate established grasp of the nature of what we are trying to measure. (K)

Few or no finding sources through FIPSE nor (at least in the California State University

System) at the state or campus level for instructional related research. There is also a near total lack of co-ordination between research in English graduate departments and schools of education. (L)

Pressure for a quick and dirty national test/indicator. (M)

1. "Political" compromise eg. avoiding controversial topics in items. (N)

2. Dilution with "pure" communication and problem solving terms. (N)

3. Failure to go for large item pool as way to handle tension between testing for improvement and accountability. (N)

1) Getting people to focus upon the improvement aspect of the goal and using assessment as a tool to that end rather than emphasizing (or getting mired in) the assessment measures and the development there of. (O)

2) Getting faculty to consider student learning in CT, PS, And CS as important and as their responsibility. (O)

This effort is too much campus-based. I think we need an indicator which addresses the political questions of return on investment, quality, and the ability of graduates to function effectively in the workplace and society. (P)

Credibility and utility at the level of faculty and students and the high probability that an initially low stakes event must become high stakes to meet shifting outcomes and political agendas. (Q)

The task is being pulled in too many directions by specialists who are too focussed on their specialties and too little concerned with the holistic conception of the project. Keep the project as an integrated whole as clearly as possible in mind. (R)

Obtaining broad school support. On the more technical level, a ...(demand?)... for oversimplification to the exclusion of the academic improvement of education. (S)

- Funding, acceptance, and the use for which employers will want to use the data (they may want GRE like scores for employment purposes.) (T)

- Cultural diversity and impact on assessment. (T)

Succeeding in getting sufficient involvement and investment by major parties for the effort to succeed--to have meaningful indicators that support change. (U)

Higher education institutions must be activity convinced that this exercise is meaningful and important. This will not happen in a design or approach that is narrowly conceived and that does not involve the higher education faculty and assessment practitioners. The national effort should fit into (and be deliberately designed to do so) existing state-level and institution-level assessment efforts in higher education. (V)

1. The focus on narrow course/discipline concerns. (Real profiles on CT,PS, & CS occur as a result of the learning process, not necessarily the curriculum. This is heresy--I know. Counseling/Advocate/Mentor function is the key--and we are missing that element in the discussion. (W)

2. Neglect of gender and ethnicity in the discussion. (W)

3. The challenge of using developmental research to inform the design of schools. (W)

4. Test mentality. (W)

1) Agreeing on the general domains of CT, PS & CS and content on "sub-specialty domains" that I do not think can be ignored. The "CT" movement was strong at the meeting--but they are not that strong in the rest of the community. 2) Developing appropriate mechanisms between states and institutions representative community that needs to have the lead on the improvement agenda. (X)

Dilution of results because of over sensitivity to politics in higher education. (Y)

How do we know (empirically) that students can't think critically? (Are CT, PS, & CS the keys to success in business--personal success or the success of business. We need baseline data. Campus resistance. Teaching to the test. Getting students to take the test. Use of results. Quibbling over definition of CT, PS, & CS. (Z)

Political - Pressure to act (AA)

Public - Explain what this means (AA)

Social - Ethnic/gender diversity - native language (AA)

Technical - Validity/reliability (AA)

Financial - Enough to do a good job. (AA)

CONCLUDING REMARKS

SAL CORRALLO: I'd like now to call on Gary, and then some brief remarks from Emerson, and that will close us out.

GARY PHILLIPS: I don't know about you, but I'm exhausted. One thing is clear to me: I think Group Two needs to go out and get drunk (Laughter.) Just a little personal true story. When I was a high school student in West Virginia, I used to skip Phys. Ed. class and go to the library and read about logic. I was eventually caught and sent to the principal for this, and the principal lectured me on the importance of physical education, whereupon I lectured him on the importance of mental education. He was not convinced, and I was punished. It isn't until today that I feel truly vindicated.

I would like to talk about "Where do we go from here?" I don't have a lot of concrete things to tell you, but perhaps enough to make it worth your while. You will be getting a copy of the Proceedings of this meeting. They will be published as an NCES publication. Also, the papers and the reviews will constitute and NCES publication, and as a matter of fact, I think we need to be entering into a phase where those are edited and corrected by you. So we'll be getting back with you about that shortly.

Also, in the very near future, there will be an internal decision paper--a working plan--that we will be developing here, which will outline what we plan to do over the next year or so. When that is available, it too will be sent to you to provide more concrete information about what we plan to do. There are some options, several scenarios that we can follow. At this point, I don't know which, and we may do all of these.

One possibility is that we will continue to work in smaller focus groups. We might take the various issues that have arisen--there have been many issue that we need to get more in-depth information about--that we need to think about. And we may convene smaller focus groups to do that.

We may also have small procurements, small contracts, to do research and development work, such as focus studies on validity and measurement issues and all of the many things we've been discussing. And we may competitively bid those, or we may find ways of doing the work--it depends on the size. If it's under $10,000, we don't have to bid, if I remember correctly.

Another thing that we might do is to create a blue ribbon panel, such as the Alexander James panel that was created several years ago to review the national assessment. If we were to do this, it would comprise high level policy people along with some staff support that would make recommendations to congress and to ourselves and others about what we should do.

Eventually, assuming that all goes well, and I don't know when this might happen, we will have a Request for Proposal. This will be the large project, to contract out the work to develop an assessment that would get this whole thing off the ground. Now the way that we do all of the surveys at NCES almost always involves a large consensus process, usually a protracted consensus process involving lots of people. This can take from six months to a year, depending on the size of the project. Also, there are always committees--technical committees, policy and advisory committees--and other groups who become involved in the process. And often there's an overall advisory committee in such projects which meets four times a year or so, to provide continued advice. We're really jot yet at the point of getting this thing off the ground, I think: still at the planning stages of this project.

We do have a budget in place. We've requested money for '92, and assuming things go well, we will have money with which we can continue the planning and development over the next year. As you know, our budget process is such that we don't yet know the budget. Each year we have to start all over again and find out what our budget will be. We do have long-term budget planning, but each year the congress must appropriate money for the projects. So in some cases we get the money, in some cases we don't, in other cases we have to rob Peter to pay Paul, to get projects off the ground.

A lot of the future of this project, I think, may depend on the work and the fate of the National Education Goals Panel, as well as the mood of the congress to appropriate money. But we do have start-up funds to continue this project through the next year, and we've requested funds for the years after that.

I believe we have heard--loud and clear--many of the issues, paradoxes, and problems we will need to deal with. I think one big issue--it is the case that we want to establish an information system--but we also want to do work that improves achievement levels of students' abilities and skills. This has many ramifications--this entire exercise--one of which is that we run into paradoxical problems in measurement when the same instrument is used to both measure progress and to improve it. This reminds me of the Heisenberg principle in physics, where we're trying to take small measurements of sub-atomic particles. In education the problem is amplified, where we are trying to measure the effects of instruction, while using the same instrument to improve and inform instruction. This problem, as a matter of fact, in the norm-reference testing world led to the Lake Woebegone effect. Where testing instruments developed for diagnostic and instructional feedback purposes were used for high stakes accountability. This led all of the states to be above the national average, as well as 90 percent of the school districts, defying the laws of mathematics. So this is a problem, but one with which we shall simply have to deal/

Another thing I want to mention to you is that in all of our projects--and particularly in this one--there is nothing we would undertake that would not happen in the public view. By that I mean that you and others will see and hear the decisions that are made, and you will participate in the process. Emerson to his credit--one of the secrets to his success at the Center--insists that all of our projects involve maximum participation by all of the stakeholders. In some cases, we may even go overboard to do that, but so far it has kept the Center in the limelight, and has improved our budget situation--though not our staffing situation yet--but Emerson's working on that.

This of course is the first step in a long process of what may be a major project at the Center. I do think it has the potential of being every bit as large and important as the national assessment. This will, of course, bring in many other groups, beyond those that are here, and we will responsibly bring them in. All the OERI-funded labs and the centers, other research organizations and centers around the country, we need to get better participation of AERA and NCME and APA, all of the interested associations and the state level organizations, ECS, MGA--I could go on and on--Council of Great City School Officers, the state organizations and institutional organizations. We will bring them all in, in time. And of course we must keep our ears to the ground on what's happening with the National Education Goals Panel, because they are a major political body that has influence on this project.

I very much appreciate what you've done. I also want to thank some people who haven't spoken, but whom you may not have seen, who have been less visible than others.. But I want to thank all of the recorders--both those who stood up to put information on the charts and those who did the recording. Pat Dabs, Sheila White, Monika Schnell, Jeff Gilmore, Chris Carr, Steve Hunt, Merrill Schwartz, Sheila Merimark, and Mary Carlson. Can we give them a hand? (Applause.)

I also want to thank Mik'al Bath and Lisa Gail who I don't think are here now, but provided excellent contract support, and were responsible for the facilities we've enjoyed here. Also, especially, I want to thank Addison Greenwood who is our own resident critical thinker, for all the work he's done. (Applause.) And most important, I want to thank Sal Corrallo and Gayle Fischer--trust me, they worked day and night tireless. You would not believe the number of hours and the number of trees these people have killed to bring about this conference. (Applause) So thank you very much,and I'll turn it over to Emerson now, for some final comments. Thank you.

EMERSON ELLIOTT: Well, my faith has been restored. I had a feeling as I listened to some of the conversations as the groups reported back this morning that it was a bit too neat and tidy, that there was a bit too much agreement. But as I heard the twenty or so folks who commented individually, my faith was restored that, in fact, this is a very difficult issue, full of all kinds of conflicts and pitfalls. Ernie, we are not ready to go out and ask someone to design an instrument. So at least that particular fear you can put aside for the moment.

NAEP goes to college, Steve, is a very catchy title. It's worth a book, at least. But I don't think we're quite ready to use it. I notice that among the groups, it was only Group One who made a reference to NAEP as possibly being a model we might follow, which I found really very intriguing. I think it's very difficult for a session that operates as this one does in a very controversial area to try to sum up. I really don't want to try to sum up, because I think there is a huge amount of information in the record, which NCES needs to assimilate. And I think that's a very important part of our task, to try to identify among the things you've said where there appears to be agreement. The many things you've said where there appears not to be agreement, and try to turn that into something that's might be appropriate for a statistical agency to do.

I had a feeling at the beginning of this morning's meeting that we really were working on the problem. That is, we weren't saying the problem didn't exist, that we shouldn't measure anything. My faith was restored on that one, as well, and before the morning was over, I head lots of people struggling with the problem, or at least it seemed that way to me.

I think we need to be very clear about the fact that there is a very strong policy push, and not delude ourselves. It's quite true, I think, that the purpose of the National Goals is to improve American education. I think the very idea of adopting the National Goals was extremely daring. It was daring because the governors and the president set themselves out to report back to the American public each year, and they knew full well that as a result the American public would keep looking at information which would keep reminding them whether or not we have made any improvement.

But it was daring for another reason: nobody had any idea how we were going to accomplish those goals, much less the objectives, of which 5.5 is one. What will it take? Well that's not clear, and that is why I think it was extremely daring for the governors and the president to adopt that. But the reason I say that we can't fight the problem is that is so glaring. As we look at what state legislatures are doing with state budgets, higher education is clearly very much affected. As we look at changes in demography that will change the demands for services at all levels of government, it is clear that higher education needs to be concerned about this issue. As we look at legislation being considered by congress--which I think has now been withdrawn, but at any rate--called ability to benefit, it raises the most profound questions about higher education and the role it plays in our society.

So I think the kinds of issues you've been grappling with here are really very up front and center issue we need to continue to work on. I think, in one area in particular--and that is improvement versus monitoring or accountability--one thing I can do is compare a little bit your conversation with the one happening right now across town, as a matter of fact, where the National Council on Educational Standards and Testing is dealing with recommendations from one of its task forces, the one on assessment. That group has made more of a distinction than the one I heard here, although maybe as I look at the full record I will discover more of it here as well, about the need for separate measures to provide separate measures to provide information that would help to improve education on the one hand, and to help to monitor education for accountability on the other. And the reason I'm so acutely aware of that is that there have been various attempts within the last year and half to use NAEP for both purposes. It has taken a great deal of work to make it clear to people that is not appropriate and won't work, in that case.

Well I think perhaps, in the preparations for this meeting and in the meeting itself, that more energy has been expended on the subject of postsecondary assessment at the federal level than has ever been expended. My conclusion to you is to say that this is the start, and only the start, of this process. Gary I think has very nicely described the set of activities that will follow on. One that I would place a bit more emphasis on is that the Center must assimilate what you have said, and try to figure out what that suggests for a way to begin the project: what really are the next steps? The next steps may well turn out to be more "researchy" than survey-like. I'd be pretty confident on the basis of what I've heard that that's fairly likely.

We will need to draw on you again. To use this as the beginning of a network that we can reach out to as we develop these ideas and extend them on to the next stage. And finally I want to thank each and every one of you very much, for devoting this energy, and coming to the session and offering your valuable comments. Thank you very much. (Applause.)

ISBN 0-16-037965-2
90000
9 780160 379659

United States
Department of Education
Washington, D.C. 20208–5653

Official Business
Penalty for Private Use, $300

Postage and Fees Paid
U.S. Department of Education
Permit No. G–17

FOURTH CLASS BOOK RATE